EFFECTS OF TASK COMPLEXITY, GLOSSING AND WORKING MEMORY ON L2 READING AND L2 LEARNING

Jookyoung Jung

University College London

Thesis submitted for the degree of Doctor of Philosophy

2017

‘I, Jookyoung Jung, confirms that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.’

Word count (exclusive of front matter, reference list and appendices):
78696

Signature _____

ABSTRACT

Over the last few decades, task-based language teaching has inspired and propelled much research into how task complexity affects second language (L2) learners' performance and development. To date, however, the task-based approach has mainly been researched in connection with learners' oral and written production, while its applicability to L2 reading has largely been unattended to. In addition, only a few studies exist that have examined the effects of glossing on L2 grammatical constructions, and so far the findings have been inconclusive. To fill these gaps, this thesis intends to examine how task complexity and glossing affect L2 learners' reading processes and comprehension, as well as their learning of target L2 constructions. Working memory capacity, which is central to both L2 reading comprehension and L2 learning, is included as a moderating variable.

The present thesis consists of three studies. The first two studies report experiments that investigated how task complexity and glossing affect L2 English reading comprehension and the learning of English unaccusative verbs and ten pseudo-word items by Korean adult learners. The participants' working memory capacity was also measured in order to examine if they moderated the effects of task complexity and glossing. The results of mixed-effects modelling revealed that task complexity and glossing had differential effects on learners' development in their knowledge of target L2 constructions, depending on the level of task manipulation (i.e., discourse level vs sentence level). L2 reading comprehension scores, however, were not influenced by either task complexity or glossing. Although working memory was found to moderate some of the relationships among the variables, no clear patterns emerged. The third study employed eye-tracking technology to explore cognitive processes during task-based reading performance and validate task-complexity manipulation. By triangulating eye-movement data with simulated recall protocols, it was found that reading processes were considerably influenced by task complexity.

ACKNOWLEDGEMENTS

I would like to express my sincerest gratitude to Dr. Andrea Révész, my supervisor, for her guidance, encouragement and inspiration throughout the entire process. This thesis would have not been possible without her committed and unflagging support. She was a marvelous mentor and a perfect role model to me in every possible way – as an advisor, a researcher, an instructor, a co-worker, and even as a working mother. I was genuinely blessed and privileged to have her as my supervisor, and I cannot possibly express in words here how much I feel grateful about her guidance.

I am also thankful to the Viva examiners, Dr. Alex Housen and Dr. Talia Isaacs, for their constructive feedback that was invaluable in revising this thesis. Their incisive comments and thoughtful suggestions were truly essential in improving the quality of this thesis conceptually as well as methodologically. I would also like to thank Dr. Amos Paran and Dr. Jackie Masterson, who reviewed the proposal of this thesis carefully and thoroughly for the Upgrade.

I also wish to express my appreciation to other professors who provided me with various forms of supports. First and foremost, Dr. Eun Sung Park's full support in data collection enabled me to carry out the studies reported in this thesis. Dr. Gareth McCray kindly shared his R-scripts, which were indispensible for analyzing eye-movement data. Dr. Frank Boers provided detailed comments to the article based on part of this thesis, which taught me how to better analyze the data and interpret the results. I am also thankful to Dr. Ronald P. Leow for his feedback on my conference presentation. Lastly,

I would like to express my gratitude to Dr. ZhaoHong Han for her insights into the relationship between L2 learning and L2 reading, which motivated this thesis.

Special thanks are also due to numerous friends and colleagues at UCL and Teachers College for their friendship and assistance. I am deeply indebted to Dr. John Rogers, who shared his brilliant works with me and taught me how to conduct mixed-effects modeling using program R. Thanks are also in order to MinJin Lee, Nektaria Kourt, Ashira Green, Hye Won Shin, and Eun Young Kang, who helped me in the course of developing research instruments and analyzing data. I am also thankful to Dr. Monika Ekiert, Dr. Jungeun Year, and Dr. Ji-Hyun Kim, who encouraged me to finish this long journey and supported me whenever I needed their insights and advice.

I would like to thank *Language Learning* and *The International Research Foundation for English Language Education (TIRF)* for their financial support to the research projects reported in this thesis.

Last but not least, my heartfelt gratitude goes to my dearest parents, Chon Ill Jung and HeaYoung Jin, my wonderful husband, Tokyou Lee, and my beautiful daughter, Seungjae Lee. Without their unconditional love, unwavering patience, innumerable sacrifices and constant encouragement, this thesis could have not been completed.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

- IL: Interlanguage
- TL: Target language
- TBLT: Task-based language teaching
- CAF: Complexity, accuracy, and fluency
- IE: Input enhancement
- GJT: Grammaticality judgment test
- WMC: Working memory capacity
- PSTM: Phonological short-term memory
- CWM: Complex working memory
- DS: Digit span
- BDS: Backward digit span
- NWS: Nonword repetition span
- OSPAN: Operation span

CHAPTER 1

INTRODUCTION

Over the past two decades, task-based language teaching (henceforth, TBLT) has garnered increasing attention from researchers as an alternative pedagogical approach in which tasks serve as a tool to engage L2 learners in language use and as the organising unit of second language (L2) instruction (Long, 2016; Robinson, 2011; Skehan, 1998; Skehan & Foster, 2001). This growing interest in the task-based approach is rooted in the assumption that tasks can function as an arena where learners can use the target language (TL) for meaningful purposes (Bygate, Skehan & Swain, 2001). In order to better assist L2 development, it has been suggested that tasks should be sequenced, from cognitively simpler to more complex ones approaching real-world demands (Robinson, 1995b, 2001a, 2011). Based on this recognition, diverse ways of classifying task characteristics and the potential consequences for the cognitive demands imposed on learners have been proposed and investigated extensively. Accordingly, various taxonomies of task characteristics have been put forward in an attempt to predict whether and how tasks with different features may have differential bearings on learners' attentional resources and, in so doing, affect their task performance as well as L2 learning (e.g., *Limited Capacity Model,* Skehan, 1998, 2009; Skehan & Foster, 2001; *Cognition Hypothesis,* Robinson, 1995b, 2001a, 2011).

As the area has matured with accumulated empirical findings, researchers have begun to conduct research syntheses and systematic reviews focusing on theoretical implications and methodological issues related to the concept of cognitive task demands (e.g., Gilabert, Manchón, & Vasylets, 2016; Jackson & Suethanapornkul, 2013; Long, 2016; Plonsky & Kim, 2016), and they have commonly identified the near-exclusive attention paid to output-based over input-based tasks. This trend in TBLT studies might be partially explained by methodological constraints. That is, production tasks allow

researchers to observe how various task features affect learners' language use fairly directly and immediately by observing learners' language production, in the form of either recorded speech or writing samples. When it comes to input-based tasks, however, how various task features affect learners' processing of aural or textual input may not be easily detectable, as the internal processes are subtle and elusive in nature and thus inherently difficult to capture. Plus, secondary evidence for task effects, such as comprehension scores or learning outcomes, may not reflect the true and immediate effects of task manipulations. In other words, it is possible that cognitive demands may have an effect at the level of input processing, which may not necessarily surface in learners' comprehension or learning scores. In a nutshell, the methodological limitations might have restricted the scope of the TBLT literature to production tasks, which should be tackled by researchers.

In addition, the effects of task manipulation on L2 development have primarily been researched in the context of interactive tasks in which input is provided by the interlocutor(s). More specifically, feedback, usually recasts, has been included as the major source of input, potentially triggering learning of a specified target linguistic construction. One explanation for this tendency is that TBLT is psycholinguistically rooted in the cognitive interactionist approach to language learning (Long, 2016), which highlights the importance of meaningful interaction, or so-called negotiation of meaning (Pica, 1987), in L2 learning. That is, task-based interaction is viewed as a useful venue for L2 practice and feedback provision that may, in turn, facilitate L2 learning. Based on this recognition, researchers have attempted to explore whether task manipulation can influence interactional patterns and thereby moderate the extent to which learners acquire a certain linguistic construction through engaging in task-based performance (e.g., Baralt, 2013; Kim, 2012; Révész, 2009).

As aptly pinpointed by Robinson (2011) and Gilabert et al. (2016), the relative ignorance regarding input-based tasks in TBLT studies may also be attributable to the lack of a theoretical framework to be called upon when explaining how differing levels of task demands might affect learners' internal processes. For example, in the case of oral production tasks, Levelt's (1989) model of speech production has served as a useful theoretical basis for predicting and explaining how task demands affect learner production, mostly in terms of complexity, accuracy and fluency (e.g., Robinson, 2005b, 2011; Skehan, 2009). In contrast, although researchers have speculated either directly or indirectly that increased task complexity might lead learners to revisit and process input more thoroughly (e.g., Robinson, 2011), no explicit theoretical proposals have been made so far to account for the influence of task demands on processing task input (Robinson, 2011).

The theoretical and pedagogical rationales for expanding the scope of TBLT research into L2 reading tasks are manifold. First of all, L2 reading is not only an important language skill most learners wish to develop, but also a major source of L2 input. Especially in the context of learning an L2 in a foreign language setting, the opportunity to engage in L2 interaction is limited, which accordingly renders the importance of L2 reading even more pronounced. Hence, for learners who are situated in a learning context with limited chances to have meaningful communication in the L2, how input-based tasks can be better utilised should be investigated further. In addition, reading is in itself a communicative activity. A prevalent misconception of reading is that the reader assumes a passive role, simply extracting information encoded in text. However, the reader takes an active part in the process of reading comprehension, such as bringing a clear purpose to the reading, activating background knowledge, and interpreting and criticizing what has been comprehended (Grabe, 2009; Khalifa & Weir, 2009). In this regard, task manipulations may have the potential to enhance the

communicative demands posed by L2 reading and encourage learners to engage in the kind of cognitive processes that are conducive to L2 learning. Furthermore, the greater accessibility of methodological tools, most notably eye-tracking technology, now allows researchers to look into learners' internal processes during reading. When triangulated with verbal reports, such as stimulated recall methodology, eye-movement data have the potential to function as a powerful research tool for documenting learners' cognitive processes when engaged in reading tasks that entail different task features (e.g., Brunfaut & McCray, 2015). Last but not least, outside the language classroom, learners should be able to perform various input-based tasks such as reading prescriptions before taking medicines, reviewing a legal contract before signing it, or enjoying literature written in the L2. That being the case, more research into how different task features affect learners' L2 reading comprehension as well as their learning of target linguistic constructions contained in texts should provide valuable insights into how to design, sequence and implement L2 reading tasks in such a way as to better assist L2 development.

With respect to L2 learning by performing reading tasks, several pedagogic approaches have been offered in order to promote the incidental acquisition of L2 features while retaining a primary focus on comprehension (Leow, 2009; Long, 1991; Sharwood Smith, 1991; VanPatten & Cadierno, 1993). Glossing (Johnson, 1982), one of the most popular textual modification techniques, is an instantiation of such a pedagogic approach. Glossing means providing information about unfamiliar linguistic items in order to aid the reader's comprehension. The underlying assumption of this technique is that it is crucial to assist learners in channelling their attention towards certain target language (TL) form-meaning mappings while processing input for meaning; otherwise they may remain unnoticed (Leow, 2015). The aim of glossing is to heighten learners' attention paid to target constructions and hence play a conducive role

in L2 development. Previous research, however, has painted a muddled picture of the role of glosses in L2 reading comprehension and L2 learning. It should also be noted that, except for a few recent studies (Guidi, 2009; Martinez-Fernández, 2010), research on glossing has been restricted to the acquisition of L2 lexis, while other constructions such as grammatical features have been largely ignored. In addition, only a few studies exist that have employed appropriate process measures, such as eye-tracking technology or stimulated recalls, to examine whether glossed items were indeed attended to by learners.

Both reading comprehension and language learning entail complex cognitive processes, such as holding linguistic information in short-term memory, comparing and matching it with existing long-term knowledge, and manipulating stored information. As working memory capacity handles these functions, it has long been researched and supported as a source of individual differences in L2 reading comprehension (e.g., Alptekin & Erçetin, 2009; Geva & Ryan, 1993; Harrington & Sawyer, 1992) and L2 development (e.g., N. Ellis & Sinclair, 1996; Miyake & Friedman, 1998; Sawyer & Ranta, 2001; Williams & Lovatt, 2003). Yet, it has not been fully investigated whether working memory capacity moderates the effects of cognitive task demands or glossing on learners' L2 reading comprehension and development of target constructions, and the findings have, overall, been inconclusive (e.g., Baralt, 2010; Kormos & Trebits, 2011).

I. The Present Thesis

In order to fill the aforementioned gaps in the literature, the current thesis attempts to explore the effects of manipulating the cognitive demands of L2 reading tasks and glossing target constructions (i.e., English unaccusative verbs and ten pseudo-word items) on reading comprehension and development in the knowledge of the targeted

constructions. In addition, eye-tracking technology was employed to document how task complexity affected reading processes and moderated the noticing of glossed items. Working memory capacity, which has long been claimed to play a central role in reading comprehension and L2 learning, was additionally examined whether and how working memory capacity moderated the effects of task complexity and glossing on L2 reading comprehension and L2 development.

II. Structure of the Thesis

The present thesis is organized as follows. The current chapter introduces the background and purpose of this research project and provides definitions of several key concepts. Chapter 2 offers an overview of the relevant literature. More specifically, the theoretical underpinnings of TBLT are introduced, followed by a review of previous empirical findings on the effects of cognitive task demands on learners' task performance and L2 development. The next section describes a model of reading comprehension process and reviews previous empirical studies on task effects on L2 reading processes and comprehension outcomes. The following section connects L2 reading and L2 learning within the SLA framework and presents an overview of research into various textual modification techniques including glossing. Then, the cognitive architecture of working memory is introduced along with previous findings on the role of working memory in L2 reading comprehension and L2 learning. Chapters 3 to 5 focus on empirical studies. Chapter 3 reports Study 1 that examined the effects of task complexity and glossing on L2 reading comprehension and development in the knowledge of the target constructions contained in texts. Study 1 also examined working memory as a potential moderator of the relationships of task complexity and glossing in L2 reading and development. Chapter 4 reports the results of Study 2, which replicated Study 1 on a larger scale and with minor modifications to the research

methodology. Chapter 5 describes Study 3, a post-hoc eye-tracking project that was designed to triangulate the findings obtained from Study 2, in an attempt to understand the underlying cognitive processes triggered by the reading tasks. More specifically, Study 3 explored the effects of task complexity on online reading processes and the noticing of glossed items as reflected in the learners' eye-movements, which was further triangulated and supplemented by follow-up stimulated recall comments. In this regard, as shown in Figure 1, the present thesis adopts a mixed-methods approach in a sequential explanatory design (Creswell & Plano Clark, 2011) at both the global (i.e., Study 2 triangulated by Study 3) and the local (i.e., eye-movement data triangulated by stimulated comments) levels. Relying on quantitative as well as qualitative data, the present thesis was able to capture not only the products but also the processes associated with reading task performance under different conditions. Chapter 6 synthesizes the findings from the three studies, discusses possible implications and limitations, and suggests directions for future research. The following section presents the definitions and operationalisation of key concepts.

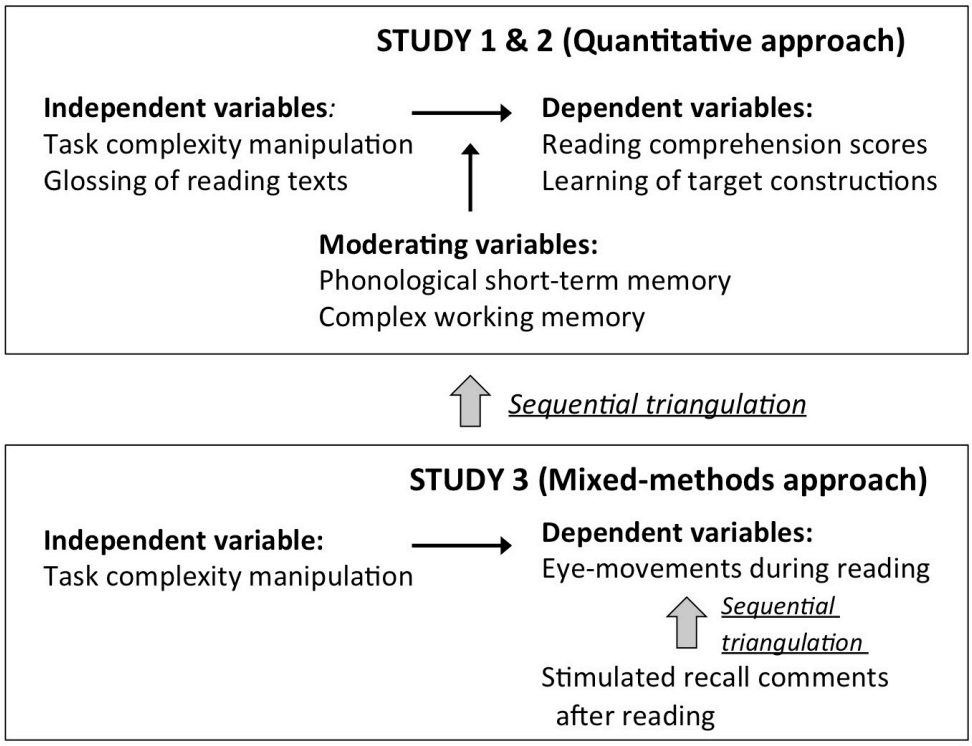

Figure 1. Research designs of Studies 1, 2, and 3

III. Key Definitions and Operationalisations

1. Input

Input is defined as linguistic samples that are available to the learner (Corder, 1967). The role of input varies across different theories of second language acquisition (SLA), ranging from Krashen's (1982) model in which input plays a necessary and sufficient role in language acquisition to the Universal Grammar framework (L. White, 1989) where input assumes a less central role. In most approaches to SLA, input is regarded as "the single most important concept of second language acquisition (SLA)" (Gass, 1997, p.1). However, it is also widely accepted that not all the input that learners are exposed is utilized for acquisition.

2. Intake

Corder (1967) coined this term, intake, defining it as what is actually internalized by the learner. More specifically, input is converted into intake when input is perceived, detected and noticed by the learner. This initial stage of language learning is often called input processing (VanPatten, 2004; Leow, 2015). Intake is the stage where new information contained in input is matched and compared with existing knowledge of the language, memory traces are formed, and generalizations and fossilizations stem from these (Gass, 1997).

3. Attention

It is a major assumption in SLA that attention plays a crucial role in language learning. According to Robinson (1995a, 1995c), the concept of attention can be encapsulated as (a) selective information processing, (b) limited capacity for information processing, and (c) the mental effort entailed in information processing. In this thesis, attention is defined as a cognitive mechanism for selection and perception of input, which is limited in capacity.

4. Awareness

Awareness is defined in different ways. For example, Tomlin and Villa (1994) define awareness as "a particular state of mind in which an individual has undergone a specific subjective experience of some cognitive content or external stimulus" (p. 193). Robinson (1995c), on the other hand, claims that awareness is a "function of the interpretation of the nature of the encoding and retrieval processes required by the task" (p. 301). In the present thesis, awareness is considered as a cognitive mechanism in which detection, registration and retrieval of input take place. Thus, awareness presupposes attention (Gass, 1997) and can be assessed as the ability to verbalize the experience or rule underlying a stimulus.

5. Noticing

Schmidt (1990) proposed two levels of awareness. The lower level of awareness is referred to as noticing (focal awareness) where stimuli are consciously registered. The higher level of awareness is termed as understanding, in which noticed information is analysed and the underlying principle, rule or pattern is recognized. According to Schmidt, noticing is a prerequisite for converting input into intake, and thus it is regarded as a crucial factor in language learning. In his later publications, Schmidt modified his original claim and embraced the importance of attention in SLA, especially with regard to the acquisition of abstract and complex rules (e.g., grammar), while viewing noticing as a facilitator of the learning process (Schmidt, 2001). That said, as pointed out by Godfroid, Boers, and Housen (2013), Schmidt's noticing can be considered as "a hybrid concept because it entails both attention and awareness" (p. 485). In the present thesis, following Godfroid et al. (2013), noticing is operationalised as focal attention, which can be gauged by an increase in eye-fixation times on target linguistic constructions.

6. Explicit vs implicit learning/knowledge

In the present thesis, explicit learning is treated as intentional learning in which learners are actively looking for patterns, often resulting in conscious knowledge (Rebuschat, 2013). By contrast, implicit learning (Reber, 1967) can be defined as a process of acquiring knowledge without intention or awareness, leading to unconscious knowledge. Various methods have been proposed to distinguish explicit and implicit knowledge from each other. For example, explicit knowledge is typically considered to be "verbalisable, learnable, declarative, and not spontaneous" (Rogers, 2016, p. 43); it can be elicited with tasks that entail controlled processing. Implicit knowledge, on the other hand, is not verbalisable, is less learnable, procedural and can be elicited with tasks that require fast and automatic processing (Rogers, 2016).

7. Task

A task can be defined as a meaning-oriented activity that requires learners to use the target language in order to achieve a specified objective (Bygate, Skehan & Swain, 2001). The key words in this definition are meaning, language use and objective. While learners may need to pay attention to both meaning and form during task completion, the primary emphasis is, by definition, on meaning, which differentiates a task from an exercise (R. Ellis, 2003). Also, a task encourages learners to function as language users in situations that mimic real-world activities. Last but not least, a task is a goal-directed activity, and thus task performance is evaluated based on the completion of the task goal, i.e., the non-linguistic outcome of the task.

8. Task complexity

Robinson defines task complexity as "the result of the attentional, memory, reasoning, and other information-processing demands imposed by the structure of the task on the language learner" (Robison, 2001a, p.28). As this definition implies,

according to Robinson, task complexity depends on proactively adjustable, task-inherent features, not on linguistic elements contained in the task. The present thesis defines task complexity as task-induced demands imposed on learners' cognitive resources while performing a task.

9. Textual modification

Textual modification entails changing textual features in order to enhance learning opportunities arising from engaging with the text (Leow, 2009). It can take various forms, such as textual simplification, textual enhancement and glossing. Textual simplification is reducing the linguistic complexity and variety of a text in an attempt to ease learners' comprehension and thereby indirectly promote L2 learning. Textual enhancement is making a particular aspect of text perceptually more salient through typographical manipulation in order to help learners to pay attention to that aspect. Glossing involves providing information about unfamiliar linguistic items in the text in order to facilitate learners' text understanding, and thereby assisting in the learning of glossed linguistic constructions.

10. Glossing

As aforementioned, a gloss is defined as information provided about an unfamiliar linguistic item in order to reduce lexical obscurity and thereby promote the level of comprehension. Various techniques can be used for glossing. In terms of language, glosses can be provided in either the first language (L1) or the L2 (e.g., Ko, 2012). Also, glosses can vary in their explicitness, ranging from a simple definition or synonym (e.g., Guidi, 2009) to an exemplar in a sample sentence containing the word (e.g., Hulstijn & Laufer, 2001). The delivery mode of glosses can also differ; both paper-based and computer-mediated glosses can be employed (e.g., Bowles, 2004). Sometimes,

multimedia, such as pictures or movie clips, are also employed for glossing (e.g., Chun & Plass, 1996).

11. Working memory capacity

Working memory can be defined as "a dedicated system that maintains and stores information in the short term, and underlies human thought processes" (Baddeley, 2003a, p. 829) and has been widely investigated as a major source of individual differences among L2 learners (Miyake & Friedman, 1998; Sawyer & Ranta, 2001). According to Baddeley and Hitch's (1974) multicomponent model, the cognitive architecture of working memory subsumes *executive control*, a limited attentional control system and two slave systems, the *phonological loop* and the *visuo-spatial sketchpad*. Among these, the executive control and the phonological loop (for more detailed operationalisations see below) have attracted particular interest from researchers in the field of first language reading (e.g., Daneman & Carpenter, 1980; Turner & Engle, 1989), second language reading (e.g., Alptekin and Erçetin, 2009; Harrington & Sawyer, 1992), artificial language learning (e.g., Martin & N. Ellis, 2012; Williams & Lovatt, 2003) and SLA (e.g., Goo, 2012; Révész, 2012a; Sagarra, 2008), among others.

12. Phonological short-term memory

In the present thesis, phonological short-term memory is considered to be synonymous with Baddeley's phonological loop. According to Baddeley and Hitch (1974), the phonological loop performs two functions. These include retaining phonological information for a few seconds during which it decays, and rehearsing and registering information stored in short-term memory. As this loop particularly pertains to the retention of sequential information, its function is typically measured with

sequences of digits or words/ nonwords that must be repeated back immediately in the order of their presentation (Baddeley, 2000).

13. Complex working memory

In the present thesis, complex working memory is treated as isomorphic to Baddeley's executive control. According to Baddeley and Hitch (1974), executive control is responsible for attentional control, conscious processing, monitoring, intentional learning and problem-solving. Complex working memory is often measured with various types of span tasks, including a reading span task (e.g., Daneman & Carpenter, 1980), a counting span task (e.g., Case, Kurland, & Goldberg, 1982) or an operation span task (e.g., Turner & Engle, 1989), which require learners to process information while retaining it in short-term memory.

CHAPTER 2

LITERATURE REVIEW

This chapter reviews theories and empirical findings that are crucial to reveal research gaps and thereby situate the present thesis within each of the associated fields. More specifically, this chapter begins with an overview of the theoretical underpinnings of task-based language teaching and previous research into how tasks affect learners' linguistic performance and L2 development. Given the centrality of the concept of task complexity, methodological suggestions are provided focusing on how to empirically validate the construct of cognitive task demands. Next, a theoretical framework for understanding the cognitive processes of L2 reading comprehension is described, followed by a review of previous studies on potential task effects on L2 reading. Then, the link between L2 reading and L2 development is briefly discussed, and various textual modification techniques, such as glossing, are introduced along with a review of previous findings on their usefulness in terms of facilitating L2 development. Finally, the cognitive architecture of working memory capacity is explained, with an appraisal of relevant empirical studies on how it may affect L2 reading comprehension and L2 development.

I. Task-Based Language Teaching

Since the 1970s, language teachers, curriculum developers, language testers and researchers in the field of SLA have discussed the need for alternatives to the traditional teacher-centred, form-oriented L2 classroom practices. In response to this need, in the 1980s, TBLT, in which tasks serve as a platform where learners can enjoy natural opportunities for meaning-oriented communication as well as a medium for infusing focus on form (R. Ellis, 2003, 2009; Long, 2016; Norris, 2009; Robinson, 2011; Skehan, 1998, 2014), was proposed as a potential approach to L2 instruction, and it has attracted

growing attention ever since. In this section, the theoretical and pedagogic bases of TBLT are introduced and empirical findings that explore task effects on learners' performance and L2 learning are reviewed, thus paving the way towards identifying gaps in the literature.

1. Task-based syllabus

Traditional L2 syllabuses are generally built around discrete linguistic units of analysis, such as words, grammatical structures, notions and functions. As learners are assumed to have the ability to integrate and synthesize isolated linguistic features one at a time in a sequential manner, this type of syllabus is called synthetic (Wilkins, 1976). Also, as the linguistic features to be taught are preselected without consideration of who the learners are and how the linguistic units will be taught, traditional syllabuses are also referred to as interventionist (R. White, 1988). The key problems of the synthetic syllabus are manifold: (a) the learner-generated sequence (so-called *built-in syllabus,* Corder, 1967) is largely ignored even though the learner's interlanguage (IL) system acts as a guiding/ constraining force en route to mastering each grammatical structure (Dulay & Burt, 1973; Pienemann, 1989); (b) learners' individual differences, such as motivation, language aptitude and cognitive capacity, are not accounted for (Dörnyei, 2009; Robinson, 1995a, 2005a; Skehan, 2002); (c) discrete target items are presented to learners in an unwarranted sequence regardless of the persistent dilemma of how to determine the complexity and learnability among linguistic features (DeKeyser, 1998; Hulstijn, 1995); (d) in a similar vein, language learning is seen as linear and additive, disregarding the dynamic relationships among linguistic features emerging as a complex adaptive system (de Bot, Lowie, & Verspoor, 2005, 2007; Larsen-Freeman, 2010, 2012); and (e) the binding power of learners' first language (L1), in terms of both the linguistic

and conceptual levels, on L2 learning is not considered (Han & Cadierno, 2010; Odlin, 1989), among others.

In response to the shortcomings of traditional synthetic syllabuses, analytic syllabuses (R. White, 1988; Wilkins, 1976) provide learners with holistic target language (TL) samples and invite learners directly or indirectly to perceive regularities in the input and acquire underlying rules through internal analysis. In addition, unlike the a priori fine-tuned nature of the synthetic syllabus, the analytic syllabus entails an a posteriori roughly-tuned pedagogic design, involving no artificial preselection or arbitrary arrangement of linguistic items. That is, as an instructional course evolves, learning objectives naturally emerge and are determined jointly by learners and teachers. Among others, the task-based syllabus (R. Ellis, 2003; Long & Crookes, 1992, 1993; Nunan, 1989, 2004) is an example of an analytic syllabus, which has been the focus of a great deal of interest over the past two decades.

The task-based syllabus entails a series of steps that begin with a careful needs analysis in order to identify real-world *target tasks* that learners will eventually do, such as reading a technical manual, giving a presentation, taking lecture notes and so forth. Once the target tasks have been identified, they are classified into *task types*, from which *pedagogic tasks* are derived. A pedagogic task, unit of a task-based syllabus, can be defined as a meaning-oriented activity that requires learners to use the TL in order to achieve a specified objective (Bygate, Skehan & Swain, 2001). As the definition implies, tasks are used not only as a vehicle for presenting appropriate TL samples to learners, but also for providing learners with opportunities for meaningful language use while attending to the TL code. In order to maximise L2 learning, individual tasks must be sequenced to match the learner's developmental level. Thus far, a number of criteria for grading and sequencing tasks have been put forward (Brindley, 1989; Candlin, 1987; Long, 1985; Nunan, 1989; Robinson, 2001a; Skehan, 1998), considering a wide variety

of task-related factors, such as task input, conditions, processes and outcomes. While this thesis will not attempt to explore each of these different proposals exhaustively, the two rival claims that highlight the performance demands of tasks from an information processing perspective, i.e., Skehan's *Limited Capacity Model* (Skehan, 1998, 2009; Skehan & Foster, 2001) and Robinson's *Cognition Hypothesis* (Robinson, 1995b, 2001a, 2011), will be examined in a later section, given their substantial impact on recent empirical studies on TBLT.

2. Task and language skills

Before advancing to an overview of the two models proposed by Skehan and Robinson, it seems important to make it clear what linguistic skills are involved in performing tasks. The definitions that have been put forward by researchers provide useful insights to clarify whether tasks involve oral or written, productive or receptive language skills, and the following examples are a few of those (e.g., Breen, 1989; Bygate, Skehan, & Swain, 2001; R. Ellis, 2003; Long, 1985; Nunan, 1989, 2004; Richard, Platt, & Weber, 1985; Skehan, 1996).

(a) A task is "a piece of work undertaken for oneself or for others, freely or for some reward. Thus, examples of tasks include painting a fence, dressing a child, filling out a form, buying a pair of shoes, making an airline reservation, borrowing a library book, taking a driving test, typing a letter, weighing a patient, sorting letters, taking a hotel reservation, writing a cheque, finding a street destination, and helping someone across a road, in other words, by 'task' is meant the hundred and one things people do in everyday life, at work, at play, and in between. 'Tasks' are the things people will tell you they do if you ask them and they are not applied linguists" (Long, 1985, p. 89).
(b) A task is "an activity or action which is carried out as the result of processing or understanding language (i.e., as a response). ... Tasks may or may not involve the production of language. As task usually requires the teacher to specify what will be regarded as successful completion of the task. The use of a variety of different kinds in language teaching is said to make teaching more communicative … since it provides a purpose for classroom activity which goes beyond practice of language for its own sake" (Richard, et al., 1986, p. 289).
(c) A task is "an activity which requires learners to use language, with emphasis on meaning, to attain an objective" (Bygate et al., 2001, p.11).
(d) A task is "a workplan that requires learners to process language pragmatically in order to achieve an outcome that can be evaluated in terms of whether the

correct or appropriate propositional content has been conveyed. To this end, it requires them to give primary attention to meaning and to make use of their own linguistic resources, although the design of the task may predispose them to choose particular forms. A task is intended to result in language use that bears a resemblance, direct or indirect, to the way language is used in the real world. Like other language activities, a task can engage productive or receptive, and oral or written skills and also various cognitive processes" (R. Ellis, 2003, p. 16).

(e) A task is "a piece of classroom work that involves learners in comprehending, manipulating, producing or interacting in the target language while their attention is focused on mobilizing their grammatical knowledge in order to express meaning, and in which the intention is to convey meaning rather than to manipulate form. The task should also have a sense of completeness, being able to stand alone as a communicative act in its own right with a beginning, a middle and an end" (Nunan, 2004, p. 4).

As illustrated above, while most definitions of task do not explicitly specify which language skills are involved, they either directly or indirectly imply that all four skills may be entailed. The definitions suggested by Long (1985) and Bygate et al. (2001) are very inclusive, implying that all sorts of linguistic activities can be involved in a task. Moreover, Richard et al. (1985) explicitly state that a task may or may not involve linguistic production, and R. Ellis (2003) and Nunan (2004) also make it clear that a task may entail productive or receptive and oral or written skills. Hence, it seems evident that a task can be directed at reading, the focus of this study, as well. However, the TBLT literature, as will be revealed later, has predominantly been geared towards learners' oral production, while other linguistic skills are relatively unattended to. As such, although the main concern of this paper is task effects on L2 reading, the following review inevitably reflects this trend in the TBLT literature.

3. Two competing models for task complexity

From an attentional capacity perspective, it is crucial to understand how various features of tasks may impose differential levels of cognitive demands, which in turn affects how learners' attention will be deployed during task completion. There are competing accounts of how this is done: Skehan's *Limited Capacity Model* (Skehan, 1998, 2009; Skehan & Foster, 2001) based on a single-resource view and Robinson's

Cognition Hypothesis (Robinson, 1995b, 2001a, 2011) based on a multiple-resource view. Although these two models are not readily applicable to L2 reading, it seems important to review them, considering the substantial body of empirical research prompted by them.

3.1. Skehan's Limited Capacity Model

Skehan, based on VanPatten's (1990) information processing view, proposes the *Limited Capacity Model*, which states that cognitively demanding tasks put pressure on learners' limited attentional resources and trigger competition among different aspects of performance for the resources that are available. In his model, the level of task demands depends on three task-related factors: (a) *code complexity*, which pertains to the linguistic complexity and variety involved in a task, (b) *cognitive complexity*, which entails processing and computational requirements, and (c) *communicative stress*, which includes time pressure, the number of participants, opportunity to control and so on. All of these factors are considered to have an important bearing on how learners' attention will be shared out during a task and how task performance will be affected in terms of linguistic complexity, accuracy and fluency (henceforth, CAF). More specifically, performing more demanding tasks will lead learners to prioritize either form (complexity and accuracy) or meaning (fluency). Additionally, within form, attention may be directed to either using challenging language (complexity) or avoiding attention-demanding structures in favour of more accurate language. Recently, drawing on Levelt's (1989, 1999) model of speech production, Skehan (2009; Skehan, Xiaoyue, Qian, & Wang, 2012) highlights the centrality of lexis, in terms of both its density and variety, as a supplementary component of CAF constructs. According to Skehan, tasks that entail the manipulation and integration of more demanding information require the conceptualization of a more complex pre-verbal message. As a result, the need to

retrieve less frequent lexical items to formulate a complex message hinders the efficient assembly of speech, thus influencing CAF.

3.2. Robinson's Cognition Hypothesis

Robinson defines task complexity as "the result of the attentional, memory, reasoning, and other information-processing demands imposed by the structure of the task on the language learner" (Robinson, 2001a, p.28). As reflected in this definition, and in contrast to Skehan's account of task demand, Robinson claims that only task-inherent features, not the linguistic elements involved, should be considered when determining task complexity. Based on a *Multiple Attentional Resources Model* drawing on Wickens's (1992, 2007) cognitive psychological model, Robinson proposes the *Cognition Hypothesis* and its associated *Triadic Componential Framework*. Within this framework, Robinson classifies task features into two dimensions, i.e., *resource-directing* and *resource-dispersing*. Along the resource-directing dimension, a task can become more demanding by increasing the number of elements involved, the amount of the reasoning required, or making reference to a displaced past time event. The cognitive and conceptual need to formulate complex content has the effect of channelling learners' attention towards lexical and grammatical encoding, which results in greater complexity and accuracy while negatively affecting fluency. By contrast, a task can also become more demanding along the resource-dispersing dimension by reducing the planning time allowed to learners or using unfamiliar task type, content or structures. In this case, learners' attention is steered towards the consolidation of, and faster access to, the existing interlanguage (IL) system, resulting in a trade-off between linguistic complexity and accuracy. In other words, in Bialystok's terms (1994), increased task complexity along the resource-directing dimension promotes the *analysis* of L2 conceptual-linguistic knowledge, whereas increased task complexity along the resource-dispersing dimension promotes *control* over L2 knowledge.

The Cognition Hypothesis further claims that increased task complexity facilitates L2 development. According to Robinson, more complex tasks lead learners to seek more help from the input provided, which results in greater depth of processing (Craik & Tulving, 1975) and long-term memory of input than is the case in simpler tasks. In addition, increased task complexity is deemed to direct learners' attention to the conceptual similarities and differences between the L1 and L2, and how these concepts are encoded linguistically in each language. Also, as for dialogic/ interactive tasks, task complexity is expected to lead to more interaction and meaning negotiation to resolve communicative breakdowns and, in turn, promote heightened attention or noticing of input and greater amount of intake/ uptake. Last but not least, Robinson also predicts that individual differences in affective factors and cognitive abilities will materialize more clearly as tasks increase in complexity.

4. Empirical studies on task complexity

Inspired by the two models above, many studies have delved into examining whether and how task manipulation affects learners' linguistic performance, interactional patterns and L2 development. This section conducts a brief review of such studies, focusing on four aspects: (a) the effects of task complexity on language production, (b) the effects of task complexity on L2 development through dialogic/ interactive tasks, (c) the relationship between task effects and individual differences, and (d) methodological issues vis-à-vis validating the construct of task complexity. This review provides the essential background in order to reveal a gap in the current TBLT literature and thereby contextualize the research focus of the current thesis.

4.1. Effects of task complexity on language production

The majority of studies in TBLT have explored how varying levels of task complexity affect learners' language production, mostly in terms of CAF measures (e.g.,

Ahmadian & Tavakoli, 2011; Foster & Tavakoli, 2009; Gilabert, 2007; Ishikawa, 2007; Iwashita, McNamara, & Elder, 2001; Kormos & Trebits, 2012; Kuiken & Vedder, 2005, 2007a, 2007b, 2011; Michel, 2011, 2013; Michel, Kuiken, & Vedder, 2007; Révész, 2011; Révész, Kourtali, & Mazgutova, in press; Robinson, 2001b, 2007; Tavakoli & Foster, 2008). Studies on monologic speech production have produced moderately converging findings. For example, Jackson and Suethanapornkul (2013) conducted a meta-analysis with nine primary studies that explored how increasing task demands affects learners' monologic oral production, and the results showed small positive effects for accuracy (d = .28) and small negative effects for fluency (d = -.16). While this global trend seems to lend support to the Cognition Hypothesis, increasing task demands was not shown to have positive effects on syntactic complexity (d = -.02), contradicting Robinson's predictions. The bidirectional task effects on accuracy and complexity, however, neatly fit into Skehan's Limited Capacity Model that is premised on a single pool of attentional resources.

When we turn to the dialogic mode, interactivity seems to moderate task effects, yielding different patterns of findings from those for monologic tasks (e.g., Gilabert, Barón, & Levkina, 2011; Michel et al., 2007; Robinson, 2007). For example, in Michel et al.'s (2007) study, it was found that increased task complexity rendered learners' speech less fluent but more accurate, with only marginal changes in linguistic complexity in the monologic mode. However, in the dialogic mode, the participants produced significantly more fluent and accurate, but structurally less complex, speech as task complexity increased. Likewise, in Gilabert et al.'s study, learners' performance changed dramatically from monologic to dialogic mode, showing strong interactivity effects. Indeed, both Skehan and Robinson predict that task effects will be materialized in distinctive ways between monologic and dialogic/ interactive tasks. According to Skehan, dialogic tasks give learners more time to regroup and replan the subsequent

message, resulting in easier lemma retrieval, and at the same time, "the presence of an interlocutor makes more salient the need to be more precise and to avoid error" (Skehan, 2009, p. 527).

Research on the effects of task complexity has also extended to the written mode (e.g., Ishikawa, 2007; Kormos & Trebits, 2012; Kuiken & Vedder, 2005, 2007a, 2007b, 2008, 2011; Révész, Kourtali, & Mazgutova, in press). An issue frequently raised here is plannability in the written mode, whereby learners are naturally allowed more time to prepare and adjust their production. For instance, in Kuiken and Vedder's (2011) study that compared task effects on L2 speaking and writing performance, linguistic complexity was affected negatively in the oral mode but positively in the written mode as task complexity increased. It appears that, in the case of written tasks, learners are able to stop their grapho-motoric progress in order to retrieve information from long-term memory or engage in a planning process, and hence linguistic and cognitive resources can be stored and used for longer, resulting in enhanced linguistic complexity of written production (Kormos, 2014). Neither Skehan nor Robinson, however, makes clear predictions about task effects on written tasks, which underscores the need to include a wider spectrum of task modes in empirical studies on task complexity (Gilabert et al., 2016).

Of particular relevance here is that task effects materialize in unique ways across different modes. Recently, task complexity was also shown to have differential effects on face-to-face and computer-mediated tasks, confirming a strong modality effect on task performance (Baralt, 2013; Heift & Rimrott, 2012; Yilmaz, 2011). Despite that, in the majority of studies, oral production tasks have enjoyed near-exclusive attention, whereas far less research has been carried out on tasks in which other language skills are involved (van den Branden, Bygate, & Norris, 2009). To be more specific, receptive language skills have not received much attention in the field of TBLT, except for a few

studies, such as Révész and Brunfaut's (2013) research into input characteristics and difficulty in listening comprehension tasks. Given that tasks are often holistic activities involving various dimensions of language use in combination (Samuda & Bygate, 2008), it seems evident that task effects should be further explored in connection with diverse language skills, such as L2 reading (García Mayo & Azkarai, 2016; Gilabert et al., 2016).

4.2. Task effects on interaction-driven L2 learning

While task effects on learners' performance have been mainly investigated in monologic mode, how manipulating task complexity affects L2 development has mostly been studied in the context of dialogic/ interactive tasks. Compared to monologic tasks, interaction-driven tasks are deemed to generate more opportunities for L2 learning, as higher communicative demands might result in more communication breakdowns and, accordingly, an increased amount of negotiation of meaning (Robinson, 2011). Motivated by this prediction, researchers have investigated whether and how task complexity affects diverse interactional features, such as negotiation of meaning (e.g., confirmation check, comprehension check and clarification request), self-repair or modified output, or language-related episodes (LREs), all of which are deemed to be conducive to L2 learning (e.g., Gilabert, Barón, & Llanes, 2009; Gurzynski-Weiss & Baralt, 2014; Kim, 2009; Kim & Taguchi, 2015; Nuevo, 2006; Révész, 2011; Révész, Sachs, & Mackey, 2011; Robinson, 2007). Findings from these studies have shown that, in general, cognitively complex tasks are likely to increase the amount of negotiation of meaning and the number of LREs (see, however, Nuevo, 2006).

There are also studies that have measured learning a specific target construction through engaging in interactive tasks, mostly employing pretest-posttest-delayed posttest designs (e.g., Baralt, 2013; Kim, 2012; Kim & Tracy-Ventura, 2011; Nuevo, 2006; Nuevo, Adams, & Ross-Feldman, 2011; Révész, 2009; Révész et al., 2014). The

target forms investigated include the English simple past (Kim & Tracy-Ventura, 2011), the English past tense and locative prepositions (Nuevo, 2006; Nuevo et al., 2011), the English past progressive (Révész, 2009; Révész et al., 2011), English question development (Kim, 2012) and the Spanish subjunctive (Baralt, 2013). The learning of target forms has typically been assessed by means of oral or written production tasks with similar designs to those of treatment tasks. In this case, obligatory contexts for using the target form are identified and it is determined whether learners produced target-like forms in those contexts. Also, additional assessment tools are sometimes employed, such as a multiple-choice receptive test (Baralt, 2010), a written metalinguistic test (Kim, 2012) or untimed grammaticality judgment tests (Nuevo, 2006; Nuevo et al., 2011; Révész, 2012), mainly in an attempt to assess development in participants' declarative knowledge of target features.

The studies above have demonstrated mixed findings. In Nuevo (2006) and Nuevo et al. (2011), the results provided very limited support for the claim that more complex tasks promote learning of the target structure (i.e., English past tense and locatives). By contrast, in Révész's (2009) study, learners who received recasts while performing a more complex task (− visual support) achieved greater gains than those who received recasts while performing a simpler task (+ visual support). Similarly, in Kim's (2012) and Kim and Tracy-Ventura's (2011) studies, where simple, + complex and ++ complex tasks were designed, learners in the ++ complex group achieved the highest gains in the development of target forms. Baralt's (2013) study also lends support to the Cognition Hypothesis, showing that performing a complex task while receiving recasts in face-to-face mode resulted in the most learning. Interestingly, in computer-mediated mode, performing a simple task led to the greatest L2 development.

As can be seen, only a limited number of studies have examined whether increasing the cognitive task demands does indeed result in greater L2 learning, and the

findings have been largely inconclusive. One reason for these mixed findings may be the fact that individual differences were not controlled in most studies. Given that learners bring various individual factors to tasks, future studies need to take these variables into account for a more nuanced understanding of the relationship between task complexity and L2 development.

4.3. Task effects and individual differences

Breen (1987) asserts that an a priori task design (in his term, 'task-as-workplan') must be tempered by what learners bring to tasks, i.e., individual differences, resulting in unique and idiosyncratic task engagement ('task-as-process'). In his Triadic Componential Framework, Robinson (1995b, 2001a, 2011) also includes learner factors under the category of *Task Difficulty* and highlights the need to investigate empirically how learners' individual differences interact with task features. He further predicts that "individual differences in ability and affective factors relevant to the cognitive demands of tasks will increasingly differentiate learners' speech production, and interaction and uptake, as tasks increase in complexity" (Robinson, 2007, p. 196). In similar vein, Norris, Bygate and van den Branden (2009) suggest that "as increasing empirical light is shed on the learner side of the equation, it is also likely that the interactions of particular learners with particular tasks will become much more predictable" (p. 245). Researchers have been keenly aware of this issue, and a variety of learner variables have been addressed in studies of task effects, such as anxiety (e.g., Kim & Tracy-Ventura, 2011; Révész, 2011; Robinson, 2007), working memory capacity (e.g., Baralt, 2010; Kormos & Trebits, 2011; Kim, Payant, & Pearson, 2015), L2 proficiency (e.g., Gilabert et al., 2011; Kim, 2009; Lai, Zhao, & Wang, 2011; Sasayama, 2016), aptitude (e.g., Kormos & Trebits, 2012), creativity (e.g., Albert, 2011; Albert & Kormos, 2004, 2011), linguistic self-confidence (Dörnyei & Kormos, 2000; Kormos & Dörnyei, 2004; Révész, 2011) and motivation (e.g., Dörnyei, 2002). So far, the findings have generally

been inconsistent, thus necessitating more research into the interface between task effects and individual differences. Given that one of the focal interests of the current thesis entails the association between learners' working memory capacity and L2 reading task complexity, studies that explored working memory capacity as a moderating variable of the effects of task complexity are reviewed here in more detail.

4.3.1. Task complexity and working memory capacity

Working memory can be defined as "a dedicated system that maintains and stores information in the short term, and underlies human thought processes" (Baddeley, 2003a, p. 829), and it has been widely investigated as a major source of individual difference among L2 learners (Kormos, 2013; Miyake & Friedman, 1998; Robinson, 2005a; Sawyer & Ranta, 2001). For instance, Kormos and Trebits (2011) investigated how learners' working memory capacity affected their performance on narrative tasks with varying levels of cognitive complexity. The participants in this study were 44 Hungarian learners of English, and their working memory was measured using the Hungarian version of a backward digit span test developed by Racsmány, Lukács and Pléh (2005). The participants also completed two narrative tasks, describing a comic strip where a storyline was given (i.e., a simple task), and making up a story with six unrelated pictures (i.e., a complex task). Task performance was analysed in terms of four global aspects, including lexical complexity, grammatical complexity, accuracy and fluency. Also, task-specific aspects were examined including accurate use of verbs, past-tense verbs and relative clauses. Overall, the participants' performances were similar across the two tasks, in terms of both global and task-specific measures, showing no statistically significant differences. Likewise, the effects of working memory capacity on learners' narrative performance were shown to be only marginal. While learners with a high backward digit span produced longer clauses overall, the ratio of subordinate clauses was similar to that of those with a low backward digit span.

Interestingly, it was learners with an average working memory span who used subordinate clauses the most frequently.

Baralt (2010) investigated the effects of task complexity and modality, i.e., face-to-face (FTF) versus computer-mediated communication (CMC), on L2 development alongside the provision of recasts. She also explored whether learners' differential working memory capacity mediated task effects on acquisition of the Spanish subjunctive. Seventy adult learners of Spanish as a foreign language were randomly assigned to one of four groups: FTF+C, FTF−C, CMC+C and CMC−C. Participants' working memory capacity was estimated using three measures: an operation span task (OSPAN), a counting span task (CSPAN) and a reading span task (RSPAN) (Conway, Kane, Bunting, Hambrick, Wilhelm, & Engle, 2005). The task used in this study was dialogic story retelling with a comic strip while receiving recasts. After completing the tasks, the participants completed an anxiety and difficulty perception questionnaire. Development was assessed using two productive tasks and one receptive task within a pretest-posttest-delayed posttest design. The results showed that + complex tasks generated the greatest gains in FTF mode but minimally so in CMC mode. Also, the − complex task in CMC mode resulted in the greatest amount of development. Learners' uptake and working memory capacity failed to predict their learning and even revealed a significantly negative relationship with development in the FTF+C group. High working memory capacity only had a significant effect on the immediate receptive test for the FTF−C group. These findings for the effects of working memory capacity were, overall, counter to expectations. Baralt suggests that adult participants in her study might have been at the peak of their cognitive ability, which could have mitigated any effects of working memory capacity being shown (i.e., a ceiling effect). Moreover, it is noteworthy that effects for both working memory capacity and task complexity were not borne out clearly in CMC mode, indicating a need for more research into how a

CMC environment mediates the effects of task complexity and individual differences in learners' task performance and L2 learning.

In sum, previous studies have produced mixed findings regarding the relationship between working memory capacity and task effects. Kormos and Trebits's (2011) study revealed that the relationship between working memory capacity and L2 narrative performance could be non-linear, and Baralt's (2010) study demonstrated that working memory capacity had an effect in face-to-face mode, but not in computer-mediated mode. Hence, more empirical investigations will be imperative to fine-tune our understanding of the role of working memory capacity in task-based language learning.

4.4. Independent measures of task complexity

As empirical research into the influence of task complexity accumulates, diverse attempts have been made in order to improve methodological rigour in this line of research, such as including a control group (Révész, 2007, 2009; Révész, et al., 2011), designing a continuous complexity scale (Kim, 2009, 2012), employing a distractor task (e.g., Nuevo et al., 2011) and comparing learners' data with native speakers' baseline data (e.g., Foster & Tavakoli, 2009; Michel, 2013). Previous studies on task complexity have, however, often eschewed how to verify the construct of task complexity, even though it is fundamental to assuring internal validity (the extent to which a causal relationship inferred in a study approximates the true relationship, minimizing the influence of extraneous variables or systematic errors) (Norris & Ortega, 2003; Révész, 2014). That is, the operationalisation of task complexity has usually been theoretically motivated, but not empirically attested independently. What researchers have typically done to estimate task complexity, if at all, is to ask learners to complete a post hoc questionnaire to elicit learners' perceptions about how challenging and difficult each task was and infer the level of cognitive demands imposed on learners (e.g., Baralt, 2013; Gilabert et al., 2009; Ishikawa, 2011; Kim, 2009; Révész, 2009; Robinson, 2001b,

2005b, 2007; Tavakoli, 2009). Indeed, according to Sasayama, Malicka, and Norris (2015), only 18 per cent of 129 studies on task complexity employed independent measures of task complexity, and 70 per cent of those used a self-reported perceived level of task difficulty or mental effort. Robinson's (2001b) questionnaire for overall perceptions of task difficulty has enjoyed popularity among researchers, perhaps due to its convenience of administration and short but comprehensive coverage of the diverse aspects of difficulty, such as stress, perceived ability, interest in the task content and motivation to complete the task. The following are the items included in Robinson's questionnaire:

(a) I thought this task was easy/ I thought this task was hard;
(b) I felt relaxed doing this task/ I felt frustrated doing this task;
(c) I didn't do well on this task/ I did well on this task;
(d) This task was not interesting/ This task was interesting;
(e) I don't want to do more tasks like this/ I want to do more tasks like this.

Learners are presented with these binary items and asked to rank each of them on a seven-to-nine point Likert scale. What has typically been done is to conduct a descriptive analysis, treating the collected scales as interval data. As pinpointed by researchers (e.g., Allen & Seaman, 2007), however, the internal consistency of self-rank data is still open to debate. Various alternative methods for assessing task complexity have been put forward, primarily motivated by *Cognitive Load Theory*, and they have been increasingly used in the field of TBLT. In a later section, some of these methods will be introduced, and it will be discussed how they have been utilized in recent TBLT research.

4.5. Gaps in the TBLT literature

The brief review of previous studies on task complexity has revealed the following concerns. First, the construct of task complexity should be empirically validated in order to enhance its internal validity and thereby contribute to the theoretical and methodological refinement of TBLT research (Révész, 2014). Second,

while research into task conditions and modalities has been continuously expanded, e.g., from monologic to dialogic, from oral to written, and from face-to-face to computer-mediated modes, task effects on receptive skills, especially L2 reading, have been largely unattended to. Provided that task effects have a distinctive influence on task performance across different modes, it seems important to expand the scope of research into L2 reading for a fuller understanding of the efficacy of TBLT. Third, considering the small number of studies and inconsistent findings regarding task effects on L2 development, it makes sense to test empirically whether increasing task complexity does indeed result in more L2 learning. Fourth, given that one of the key postulations of TBLT is that task effects are implicated in the process of linguistic encoding (Kormos, 2011; Levelt, 1989) and promote attention to TL construction(s) (Robinson, 1995b, 2001a, 2005b; Schmidt, 1995), more research is required to examine whether these theoretically presumed linguistic and cognitive processes do indeed occur during task performance (Révész, 2014). Lastly, how individual differences moderate task effects should be investigated in order to draw more readily applicable pedagogical implications. As Robinson (2011) suggests, systematic research into the interplay between task effects and individual differences will provide valuable information about how to better match learners to instructional conditions and practice sequences. Hence, the next sections will conduct an overview of the literature relating to the issues addressed here, namely, (a) methodological concerns regarding assessment of the level of task complexity, (b) the potential influence of task manipulation on L2 reading, (c) learning L2 form-meaning connections through reading, (d) various process measures for documenting learners' internal processes during task performance, and (e) the role of working memory capacity in L2 reading and L2 learning.

II. How to Assess Task Complexity

As previously pointed out, researchers have increasingly recognized the need for empirical validation of task complexity (Norris & Ortega, 2003, 2009; Révész, 2014). *Cognitive load theory* (de Jong, 2010; Paas & van Merriënboer, 1994a; Sweller, van Merriënboer, & Paas, 1998), which is premised on a cognitive architecture consisting of a limited working memory capacity, can provide a useful framework to pursue this. Cognitive load theory is concerned with how to design instructional tasks that use learners' limited cognitive resources efficiently so that they can better apply and transfer acquired knowledge to new situations. Cognitive load theory and TBLT share many commonalities, in that both are instructional frameworks seeking to maximize learning opportunities through developing more efficient teaching materials, and the concept of cognitive demands is deemed to be key in attaining this goal. To be more specific, the crucial factor in cognitive load theory is the concept of *cognitive load*, which is "not simply a by-product of the learning process but as the major factor that determines the success of an instructional intervention" (Pass, Tuovinen, Tabbers, & van Gerven, 2003, p. 64). For the theoretical concept of cognitive load to be functionally useful, it is considered necessary to assess cognitive load in a more direct and objective fashion (Brünken, Plass, & Leutner, 2003). That is, through repeated comparisons of the predicted cognitive load of an instructional design and empirical assessment of the actual cognitive load experienced during task performance, it becomes increasingly viable to predict the level of cognitive load in an early design stage. Thus, cognitive load theory seeks empirical accounts of how the actual level of cognitive load relates to learners' performance and learning.

Brünken et al. (2003) provide a useful classification of methods for measuring cognitive load, as presented in Table 1. The objectivity dimension pertains to whether data derive from learners' self-reports or objective observations of learners' behaviour

or performance. The causal relation dimension categorizes methods based on the relationship between an observed phenomenon and an actual attribute, i.e., cognitive load.

Table 1. Classification of methods for measuring cognitive load (*Source*: Adopted from Brünken, Plass, & Leutner, 2003, p. 55)

	Causal Relationship	
Objectivity	Indirect	Direct
Subjective	Self-reported invested mental effort	Self-reported stress level Self-reported difficulty of materials
Objective	Performance outcome measures Physiological measures	Brain activity measures Dual-task performance

1. Direct and indirect subjective methods

Subjective methods involve rating scale techniques using post-treatment questionnaires in which learners report the amount of mental effort they have invested or the level of fatigue, stress or difficulty they felt while completing a given instructional task (Paas & van Merriënboer, 1994b). The rationale for using this method is twofold. First, learners are able to interpret cognitive load scales designed by the researcher; second, learners can translate their mental effort invested retrospectively (Cierniak, Scheiter, & Gerjets, 2009). Paas, van Merriënboer, and Adams (1994) demonstrated that self-reported mental effort may be more sensitive to variations in cognitive load than obtrusive and laborious physiological measures, and their rating scales have been adopted in many studies (e.g., Ayres, 2006; Kalyuga, Chandler, & Sweller, 1999; Van Gerven, Paas, van Merriënboer, & Schmidt, 2002). Paas et al.'s (2003) meta-analysis of studies that measured cognitive load also revealed that self-ratings have remained more popular than more objective methods among researchers, because they are easy to administer, unobtrusive and inexpensive. As mentioned in an earlier part of this paper, in TBLT studies, questionnaires asking about learners' perceived difficulties have also been preferred as a way of inferring the level of cognitive complexity of the tasks used (e.g., Baralt, 2013; Gilabert et al., 2009; Kim &

Tracy-Ventura, 2011; Révész, 2009, 2011; Révész, Hama, & Sachs, 2014; Révész, Michel, & Gilabert, 2016; Robinson, 2001b, 2007; Sasayama, 2016). It should be noted, however, that a number of methodological issues remain unresolved, as pinpointed by de Jong (2010): (a) it is not clear if learners can estimate the average of demands that constantly fluctuate during task performance; (b) in the same line of logic, the temporal variants in cognitive demands cannot be captured as learners recall the level of demands as a whole, a posteriori; (c) the meanings of words and scales are inherently susceptible to subjective interpretation, undermining internal consistency; and (d) variations among questionnaires are likely to magnify inconsistent findings, which hinders researchers attaining convergent validity (the degree to which multiple measures of constructs that are theoretically assumed to be related are in fact related).

2. Indirect and objective methods

Learners' performance or knowledge acquisition scores can also be used as a measure of cognitive load. As cognitive load is inferred from learners' observable performance or learning scores, this method is indirect and objective. A typical procedure is to design two or more variants of instruction with the same material, based on the assumption that the intrinsic load induced by the material is the same. Thus, the better performance or learning outcome the learner exhibits, the less cognitive load is induced by the instruction. Yet, it should be noted that learners' performance and learning can be affected not only by different types of instruction but also by the measurement method, and individual differences may also come into play (Brünken et al., 2003).

3. Direct and objective methods

Arguably, direct and objective methods may generate the most accurate estimate of the cognitive load imposed on learners. Measuring brain activity is an example of

such methods. For example, functional magnetic resonance imaging (fMRI) technique has been used to capture brain activity during task performance (e.g., E. Smith & Jonides, 1997). Also, Murata (2005) used a wavelet transform of electroencephalographic (EEG) signals to measure cognitive load, and it was found to be a sensitive indicator of levels of cognitive load with high precision. Yet, the connection between cognitive load and brain activity is not yet fully documented and is in need of further empirical validation.

Another more widely used technique is a dual-task methodology. The rationale is that if a learner has to simultaneously carry out two different tasks, calling on the same cognitive resources, the resources available have to be distributed between tasks, and this competition can be evidenced in performance on the secondary task. Brünken et al. (2003) describe the conditions for an ideal secondary task to carry out dual-task analysis. First, the secondary task should require the same cognitive resources as the primary task, so that the secondary task is dependent on primary task performance; second, performance on the secondary task should be reliably and validly assessed; third, the secondary task should only moderately interfere with primary task performance, not to the extent that it suppresses simultaneous task performance; and fourth, the secondary task should consume the available cognitive resources flexibly so that learners can keep performing the instructional task while responding to the secondary task's requirements (Brünken et al., 2003).

Subjective time estimation is often referred to as a useful method to measure cognitive load. Learners are usually asked to judge the time taken for task completion in the absence of an external timing device, and it has consistently been shown that estimated time duration becomes less accurate as the cognitive load of the task increases (Block, Hancock, & Zakay, 2010; Hicks, Miller, & Kinsbourne, 1976). Thomas and Weaver (1975) provided the theoretical basis for this method. That is, as nontemporal

task demands increase, less attention is left for processing temporal information, and as a result estimated time duration becomes more inaccurate. More specifically, under the prospective paradigm where participants are aware of the upcoming time estimation to be made at the outset of the task, it has consistently been found that estimated-to-real duration ratio decreases with increasing cognitive load. By contrast, under the retrospective paradigm where participants are unaware of the subjective time estimation task until it has to be made, the estimated-to-real duration ratio increases with cognitive load (Fink & Neubauer, 2001).

Measuring reaction time to visual or auditory stimulus is another example of the dual-task method. When using this method, learners are asked to react to a specific signal as soon as possible while performing a primary task. While the reaction task consumes few cognitive resources and thus does not interfere with the primary task, responding to the stimulus temporarily depletes the available resources. Thus, this method "minimizes the interference between the two tasks and maximizes the exhaustion of the free capacity" (Brünken et al., 2003, p. 57). Given these merits, reaction time measures have widely been used in cognitive load theory studies (e.g., auditory stimulus in Brünken et al., 2003; visual stimulus in Cierniak et al., 2009).

4. Application in TBLT studies

Some of the methods reviewed above have recently been employed in TBLT studies in an attempt to test the level of task complexity independently. For instance, Kim et al. (2015) employed stimulated recalls and interview protocols in addition to self-reported perceptions of task difficulty in order to validate the level of task complexity. The analysis of recall protocols revealed that a task designed to be more complex indeed triggered more comparisons and evaluations of task components. Perceived level of task difficulty, however, failed to discriminate participants who performed simple versus complex tasks. Also, Baralt (2013) employed not only a

perception questionnaire but also a retrospective time judgment task as an additional source for estimating cognitive complexity. In her study, learners were asked to estimate how long they believed it took them to complete each task, postulating that the greater the demands imposed on the learner, the more time he or she would judge had passed (Paas, Tuovinen, Tabbers, & van Gerven, 2003). The findings of this study showed that the retrospective time judgment measure matched Baralt's operationalisation of task complexity, while a perception questionnaire failed to do so. More specifically, participants who performed the complex version estimated the time taken for task completion to be significantly longer than the time actually taken. It should be noted, however, that Baralt analysed subtracted values (the difference between estimated and real time duration), not estimated-to-real duration ratios, which could have lowered the internal validity of the results.

Researchers seem to agree that multiple sources of evidence are desirable for a more robust validation of the constructs of task complexity (Révész, Hama, & Sachs, 2014; Révész, Michel, & Gilabert, 2016; Sasayama, 2016). For example, Révész et al. (2014) used three different methods, i.e., expert judgments, eye-tracking technology and dual-task methodology, in order to independently assess the validity of cognitive task complexity manipulation. First, two experts were invited to respond to 5-point Likert scale survey questions to rate the cognitive complexity of tasks. Also, eye-tracking technology was used, based on the assumption that the greater the number and duration of eye-fixations, the more demanding tasks were. In addition, the participants were asked to react to a colour change in the computer screen background, a slower and less accurate reaction indicated that a task was cognitively more demanding. The analysis of data obtained from these three validation methods revealed that the manipulation of task complexity was, overall, successful. In a more recent study, Révész et al. (2016) compared the validity of dual-task methodology, self-ratings and expert judgments. In

this study, the participants carried out simple and complex versions of three oral tasks. Half of them completed the tasks under a dual-task condition. The other half, on the other hand, performed the tasks under a single-task condition but provided self-reports on the perceived level of task difficulty. ESL teachers were invited to rate the anticipated level of mental effort required for, and task difficulty of, each task. The analysis of these three sources of data confirmed that complex tasks were indeed perceived and rated more demanding by the participants and teachers, supporting the validity of the methods for assessing task complexity. Similarly, Sasayama (2016) utilized self-reported perceptions of task difficulty, prospective and subjective time estimation, and dual-task methodology. She found that only large differences in the number of task elements were detectable, while smaller differences did not make a significant change to the level of cognitive complexity. Based on this finding, Sasayama underscores the importance of independent measures of task complexity in order to attest to whether designed task features exercise putative effects on task complexity.

5. Summary

This section has covered diverse methods that have been utilized in cognitive load theory studies, such as self-reports, physiological measures and dual-task methodology, which have enabled researchers to assess the actual cognitive load put on learners' mental resources. Also, several recent TBLT studies that have utilized these methods have been reviewed, casting light on their potential usefulness for validating task complexity. As demonstrated above, borrowing the methods used in cognitive load theory studies seems useful to enhance the validity of research into task effects and thereby refine the theoretical underpinnings of TBLT.

III. How Task Affects L2 Reading

As pinpointed in the earlier section, both Skehan's Limited Capacity Model and Robinson's Cognition Hypothesis predict and explain task effects on speech production, and hence they are not directly applicable to L2 reading tasks. Given that understanding the processes and components involved in L2 reading is essential in order to better understand how tasks may influence L2 reading, Khalifa and Weir's (2009) cognitive processing model for reading comprehension was considered as an ideal theoretical basis of the present thesis. Unlike previous models of reading (e.g., Kintsch & van Dijk, 1978; Perfetti, 1999; Rayner & Pollatsek, 1989; Stanovich, 1980) whose scope was restricted to the cognitive processes of text understanding, Khalifa and Weir's framework incorporates and stresses the role of metacognitive mechanism in reading comprehension. The highlighted role of metacognitive function allows the model to account for the influence of task demands on reading, by viewing reading as a cognitive process that constantly reacts to the goal of the reading task. Given that task demands was the focus of this study, this emphasis on metacognition fucntion made this model particularly suitable for the purposes of this thesis. The description of Khalifa and Weir's reading process model is followed by a review of previous research into task effects on L2 reading.

1. Cognitive processing model for reading comprehension

As displayed in Figure 2, the model presupposes three sources of knowledge: metacognitive activity, the central core and the knowledge base. The knowledge base is what learners bring to the reading task, such as general knowledge of the world, text-related knowledge (e.g., text structure, genre and topic), and linguistic knowledge (e.g., orthography, phonology, lexical knowledge and syntactic knowledge). A solid knowledge base can facilitate reading comprehension. For example, if a reader has highly developed linguistic knowledge (e.g., advanced or native-like proficiency) as

well as sound background knowledge related to the topic and genre of the text, the resulting comprehension is likely to be accurate and robust. By contrast, if a reader lacks the required knowledge base to process the text, the reader may resort to other sources of knowledge in order to compensate for this deficiency.

The central processing core entails cognitive processes of reading that begin with word recognition, followed by lexical access, syntactic parsing and creating meaning propositions at the clausal or sentence level (i.e., micro-structures). Word recognition can be defined as "the perceptual process of identifying the letters and words in a text" (Field, 2004, p. 234). Automatized word recognition is regarded as a prerequisite for fluent reading comprehension (e.g., Gorsuch & Taguchi, 2010). Word recognition also involves activating relevant semantic and syntactic information such as the word class and grammatical structure of a lexical item, mainly through an automatic spreading activation mechanism (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). That is, recognition of a word form automatically activates neighbouring semantic information, such as collocates or similar meanings in the lexical network, and syntactic information, such as the morphological function and syntactic regularity of a word (Kieffer & Lesaux, 2012; Shiotsu, 2009). As efficient word recognition requires knowledge of orthographic and phonological regularity and sufficient sight vocabulary size through extensive exposure to texts, slow and inefficient word recognition generally becomes the first obstacle for most beginning L2 readers.

As readers recognize words, almost simultaneous syntactic parsing takes place, using morphological and structural information taken from the words, in order to integrate the words into phrasal and clausal meaning units (Fender, 2001). Syntactic parsing involves various processes, such as disambiguation (suppressing alternative meanings), tracking referents and default processing strategies (e.g., subject-verb-object ordering expectations, preferences for certain clause structures and repair strategies), to

name a few (Grabe, 2009). Encoded semantic propositions serve as building blocks for text comprehension as a whole (Fender, 2001). Word recognition, lexical access and syntactic parsing can be viewed as lower-level processes, which draw on linguistic knowledge.

The propositional units produced by lower-level processes are assembled to create a mental model of the text as a whole (i.e., macro-structure). When connecting propositions, 'bridging' inferences are used to make sense of intra-textual vagueness and maintain coherent relationships between propositions (Kintsch, 1998; Pressley, 2006). Among propositions, the ones that are most strongly activated and share multiple networks with other propositions become the main ideas of the text. At the next stage of text-model building, a broader discourse-level structure emerges. Skilled readers are experienced in identifying the hierarchical structure of an entire text and determining its central meaning. Finally, a text-level representation may be linked to other related texts if necessary, following transformational macro-rules of deletion, generalization and integration. For example, when writing an essay after reading multiple articles, intertextual representation is emphasized.

Metacognitive activity, the left-hand column in Figure 2, has particular relevance to studies investigating task effects on L2 reading, as it involves setting goals, monitoring and remediating text understanding where necessary. The goal-setter determines the type of reading comprehension that should be aimed for and the speed and scope of reading required. More specifically, according to the purpose of reading, the reader engages in either *careful* or *expeditious* reading, which takes place at either *local* or *global* level. Local comprehension refers to extracting propositions at the level of micro-structure, such as a clause or sentence. Local comprehension is strongly associated with linguistic knowledge, requiring lexical access, syntactic parsing and micro-level proposition encoding for understanding explicit text-based information

(Alderson, 2000; Cohen & Upton, 2006). Global comprehension entails understanding the structure of a text as a whole, building macro-propositions beyond the level of micro-structure (Kintsch & van Dijk, 1978).

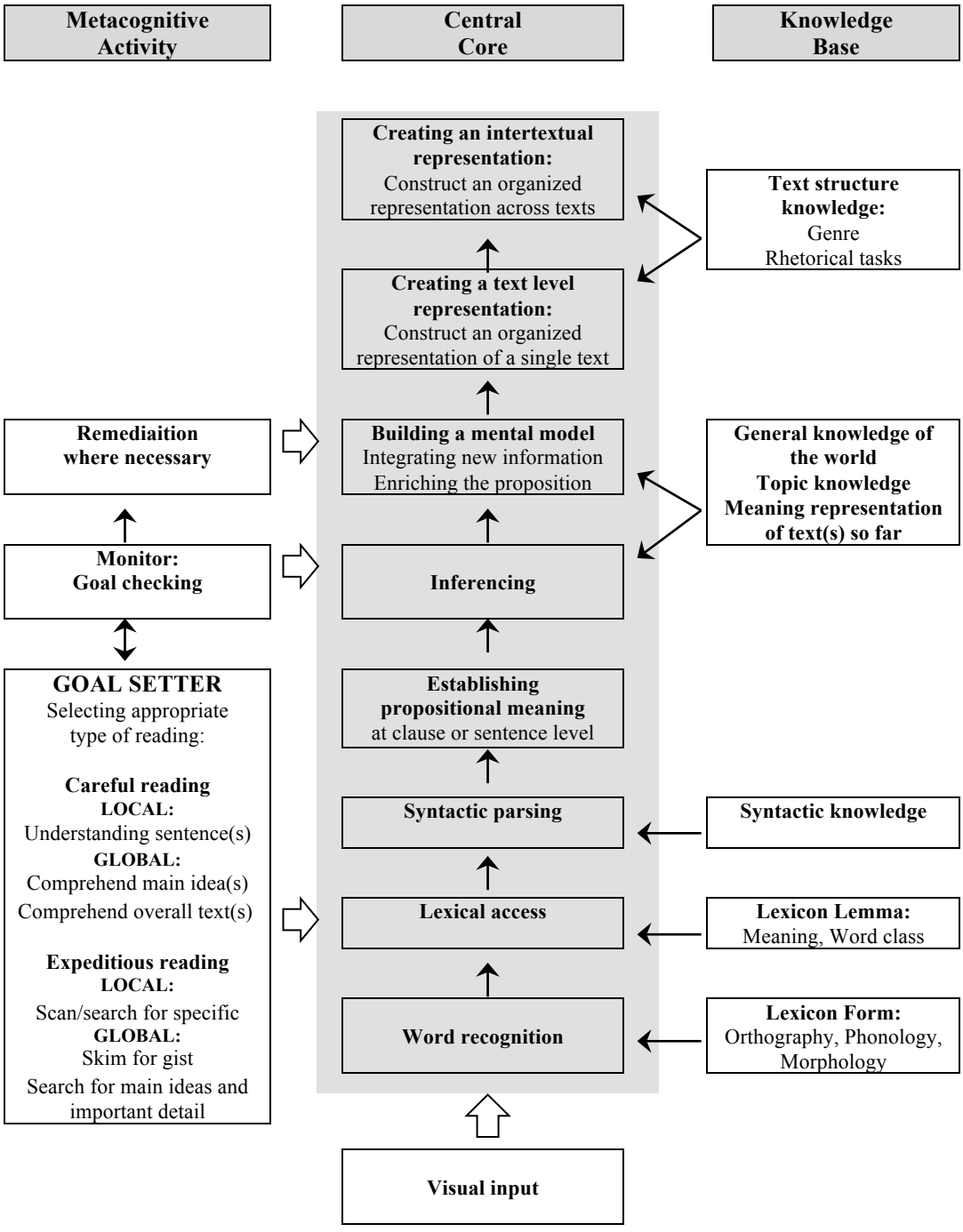

Figure 2. Cognitive processing model for reading comprehension (*Source*: Adopted from Khalifa & Weir, 2009, p. 43)

Careful reading is what researchers typically have in mind when investigating reading, and thus extensive research on reading has focused on careful reading. Hoover

and Tunmer (1993), for instance, specified that they focused on "comprehension that is intended to extract complete meanings from presented material as opposed to comprehension aimed at only extracting main ideas, skimming, or searching for particular details" (p. 8). Careful reading may take place at local or global level and is usually based on slow, attentive, sequential and incremental reading for comprehension (Khalifa & Weir, 2009). By contrast, expeditious reading has been largely ignored, even though researchers generally agree that this type of reading can impose even greater problems for inexperienced L2 readers. Expeditious reading can be challenging because it requires rapid word recognition and proficient syntactic parsing, which depends on automatized lower-level processing skills built through sufficient practice of reading in the TL. Unlike careful reading, expeditious reading includes quick, selective and efficient reading (Khalifa & Weir, 2009), often to achieve a particular goal of reading as in the case of skimming, search-reading or scanning. Skimming is reading to obtain the gist, general impression or overall discourse structure of a text. Search-reading involves locating specific information on a predetermined topic in advance of reading, and scanning is reading selectively to find specific words, phrases, numbers and dates, usually at the local level. According to Khalifa and Weir (2009), readers choose various permutations of the two dimensions of reading (i.e., local vs global and careful vs expeditious) in such a way that the goal of reading is most likely to be met.

The role of monitoring is contingent on the type of reading pursued, and thus monitoring functions in accordance with the goal of reading. Monitoring occurs in all stages of reading, from checking word recognition to evaluating the text-level representation and extracting the writer's intentions and text structure. Unskilled readers are less competent at self-monitoring and checking the meaning representation for consistency, whereas skilled readers are capable of regulating and adjusting their reading behaviour according to the different purposes of reading (e.g., Horiba, 2000,

2013). In this regard, the goal-setter and monitor serve as metacognitive mechanisms that enable readers to call upon different reading strategies and skills with different reading goals in mind.

When Khalifa and Weir's model is applied to TBLT, the following theoretical assumptions can be drawn for analysing and manipulating the cognitive demands of L2 reading tasks. First, if a reading task can only be accomplished through careful and thorough processing of a text, the task is expected to be more demanding than a reading task that can be carried out through superficial reading of the same text. Next, if a reading task is to be completed expeditiously within a time limit, it is likely to be more cognitively challenging for L2 readers than performing the same task without such time pressure. Lastly, when these two dimensions (i.e., depth and speed of reading) are held constant, a reading task that entails a wider scope of reading (e.g., multi-paragraph essays) would be more complex than one that can be carried out by processing only a limited scope of reading (e.g., a list of words). These hypotheses need to be addressed in future studies.

2. Previous studies on task effects on L2 reading

Few studies have investigated how the same person performs differently when reading for various purposes. Nonetheless, some studies have demonstrated that readers might change their way of reading according to the situation (Horiba, 2000, 2013; Taillefer, 1996; Yoshimura, 2006). For example, Taillefer's (1996) study showed that L2 readers might rely on their L2 knowledge to a differential extent in reading tasks with varying levels of complexity. Fifty-three native French speakers read English texts for two different purposes: reading to prepare for an upcoming debate (i.e., receptive reading) and reading to locate occurrences of keywords (i.e., scanning). Taillefer assumed that receptive reading is more complex than scanning, as scanning is by and large a simple cognitive matching task, searching for what is sought and what is already

given. However, receptive reading necessitates not only word recognition skills, but also syntactic parsing for integrating words, constructing clausal-meaning units, holding them in short-term memory while processing subsequent sentences, and building a coherent text model. Thus, receptive reading is considered a more complex process than scanning, in terms of both cognitive and linguistic demands. The participants' reading comprehension was measured with cloze tests, and L2 English proficiency was assessed with TOEFL. The results of multiple regression analyses indicated that the amount of variance in L2 reading comprehension accounted for by L2 proficiency decreased considerably in a less complex reading task (i.e., scanning) compared to a more complex task (i.e., receptive reading). Based on the findings, Taillefer suggested that L2 proficiency might make a more marked contribution to a more complex L2 reading task.

Horiba (2000) investigated how L2 readers' control over their reading process might differ from that of L1 readers across different types of texts and tasks. This study consisted of two experiments, Experiment 1, focusing on the effects of text type, and Experiment 2, focusing on the effects of task type. The first experiment included seven native and seven nonnative readers of Japanese. Reading materials were two newspaper essays and two short folktales. The participants read the texts as they normally would, while verbalizing their processes in the language they felt more comfortable. After reading, they produced oral summary recalls for the essays and written free recalls for the stories. The findings of this experiment showed that the nonnative readers had to engage predominantly in processing textual information, and thus had few resources left for regulating their reading processes according to the different types of texts.

Experiment 2 delved into task-induced effects on native and nonnative readers' processing. Participants in this experiment were fourteen native and fourteen nonnative readers of Japanese, and they were assigned to either a read-freely condition or a read-for-coherence condition. The reading materials and the procedures were the same as

those for Experiment 1. The results revealed that task type did not affect their summary recall products. It was also found that the native readers were competent enough to regulate their reading processes according to the task type, while maintaining a sound understanding of the content of the texts. By contrast, non-native speakers tended to use more relational and integrative processing in the read-for-coherence condition than in the read-freely condition. It was also found that a large proportion of their cognitive resources was allocated to basic text-based, lower-level processes. In sum, in this study, native readers were competent and flexible in controlling their own processing and allocating their cognitive resources strategically according to the type of text and the task, whereas nonnative readers were not able to control their reading processes, mainly due to linguistic demands.

More recently, Horiba (2013) again investigated how task instructions affect L2 readers' text processing and comprehension based on the assumption that strategic and flexible text processing is an important factor of successful comprehension and knowledge acquisition. In Experiment 1, 84 L1 Japanese participants were instructed to read argumentative essays in either Japanese or English in one of the following conditions: reading to understand expressions, reading for image and reading for critique. The level of comprehension was measured via L1 free written recall. A two-way ANOVA revealed that both L1 and L2 text comprehension, in terms of the amount of content recalled, did not differ across the different reading situations. In Experiment 2, 28 participants were provided with the same texts and the same task instructions, but instructed to do think-aloud protocols while performing the task. The results revealed that the process of L2 text comprehension differed when readers processed a text for different reading goals. More specifically, when reading to understand expressions, the participants allocated greater amounts of mental resources to lower-level processes, paying more attention to unfamiliar words or phrases. But when reading for critique, it

was found that the participants utilized more resources for higher-level processes in order to interpret their text understanding and the author's intention. When reading to visualize the content, the pattern of mental resource allocation was characterized somewhere between reading for expressions and reading for critique. The findings led Horiba to conclude that task effects on L2 reading might be implicated at the level of text processing, rather than comprehension outcome.

When the issue turns to L2 learning from engaging in different reading tasks, Yoshimura's (2006) study provides some insights. Motivated by Izumi's research (Izumi, 2002, 2003; Izumi, Bigelow, Fujiwara, & Fearnow, 1999) in which the Output Hypothesis (Swain, 1985, 1995; Swain & Lapkin, 1998) was tested, Yoshimura investigated whether manipulating foreknowledge of a post-reading task led to different reading behaviour, text comprehension and noticing of L2 features. The participants were 57 Japanese university students learning English as a foreign language. They were randomly assigned to one of three groups that had different tasks to perform after reading a text for five minutes. Post-reading tasks were reading for memorization, reading for retelling and reading for visualization (no output). After reading the text, however, the post-reading tasks were not administered. Rather, the participants were asked to (a) report their reading behaviour by completing a retrospective questionnaire, (b) answer true-or-false comprehension check questions, and (c) fill in the blanks in the text with appropriate verbs. The results showed that the output groups (i.e., reading for memorization and reading for retelling) used more diverse reading strategies, such as translating into their L1, matching their existing L2 knowledge with target constructions in the text, and paying more attention to forms in the texts than the visualization group. Also, the participants in the memorization group reported more use of L1 translation and monitoring strategies than those in the retelling group. It was further revealed that scores on the verb production test were higher for the memorization group, lower for the

retelling group and the lowest for the visualization group. By contrast, comprehension scores were not significantly different across groups. From the findings, Yoshimura suggested that learners' reading processes could be affected by foreknowledge of the required task output, and that different task instructions might have a differential impact on language processing.

More importantly, Yoshimura's (2006) study demonstrates that it is viable to use tasks to promote learners' reading for acquisition without interrupting reading for comprehension. That is, if the target construction is regarded as essential for task completion, learners may pay more attention to TL features in the text during reading. However, there are also limitations to this study. First, spending five minutes for reading an 81-word long text could have been too long, and thus might have mitigated the observed variation contributed by different output tasks to comprehension. In addition, Yoshimura interpreted the participants' verb production in a fill-in-the-blanks test as an indicator of noticing. Yet, given that they were given five minutes to read a very short text while planning to retell or memorize it, it seems highly likely that the verb forms were processed at a higher level of awareness accompanied by metacognitive efforts. It should also be noted that this study included no concurrent measures for documenting learners' on-line reading processes. Hence, whether the foreknowledge of the post-reading task did indeed promote noticing of TL in the text can be only speculative at this stage. That said, more sensitive tools with clearer operationalisation of noticing, such as verbal reports or eye-movement data, could paint a more precise picture of L2 processes across different types of L2 reading tasks.

3. Summary

This brief overview of cognitive processing models for reading comprehension (Khalifa & Weir, 2009) has demonstrated that reading entails complex and interactive processes, in which readers play an active role in communicating with a text with

various purposes in mind. According to the purpose of reading, the goal-setter modulates the entire reading process so that various sub-skills come into play, by making metacognitive decisions about whether to read carefully or expeditiously in order to achieve a local or global understanding of the text. Going one step further, theoretical suggestions on how to apply Khalifa and Weir's model to analyse and manipulate L2 reading tasks were proposed. The empirical studies reviewed here show that learners might activate different reading processes in tasks that entail different goals, instructions and output formats (Horiba, 2000, 2013; Taillefer, 1996; Yoshimura, 2006). Yoshimura's study, in particular, shed light on the viability of manipulating L2 reading tasks to facilitate L2 learning without interrupting text comprehension. As such, in the next section, theoretical underpinnings relevant to the relationship between L2 acquisition and input comprehension are discussed, followed by a review of empirical studies on various text modification techniques designed to promote L2 learning from L2 reading.

IV. L2 Learning from L2 Reading

L2 reading is an important language skill that most L2 learners seek to develop and, at the same time, a means to acquire the L2, serving as a major source of comprehensible input (Eskey, 2005; Krashen, 1982). Certainly, these two aspects of L2 reading are not mutually exclusive but rather somewhat interconnected, presumably sharing a symbiotic relationship. That is, greater knowledge of L2 enables more proficient text processing for better L2 reading comprehension, which in turn results in further growth in L2 competence. As such, L2 reading instruction should be designed and implemented aiming to achieve these dual goals, i.e., enhancement of comprehension ability and development in L2 competence, in a way that can be supported by theoretical and empirical research.

1. Tension between comprehension and acquisition

So far, diverse views have been suggested for the role of reading in L2 learning with varying pedagogical foci. In particular, two major approaches to L2 reading have emerged: the literacy-based approach, which views the development of L2 reading ability as a comprehension skill, and the acquisition approach, which focuses on the development of L2 competence through L2 reading. As noted by many researchers (e.g., Bernhardt, 2005; Grabe, 2005, 2009; Han, Anderson, & Freeman, 2009; Pulido, 2007, 2009; Urquhart & Weir, 1998), however, these two approaches have followed diverging paths showing little overlap. As a consequence, the current picture of the relationship between L2 reading and second language acquisition (SLA) is far from clear.

It is not that there has been no attempt to combine the literacy-based approach and the acquisition approach. For one, Krashen (1982) proposed the *Input Hypothesis,* arguing that free voluntary reading for comprehension automatically serves the purpose of acquiring an L2. Yet, this argument has been challenged by many researchers on an empirical basis (e.g., Izumi, 2002, 2003; Sharwood Smith, 1986; Swain, 1985; Swain & Lapkin, 1998; VanPatten, 1996, 2004). That is, unlike L1 learners' implicit language learning, L2 learners are generally inclined to process TL input solely for the sake of comprehension, which does not necessarily result in restructuring their interlanguage (IL) system. L2 learners' difficulty in processing input for both meaning and form (i.e., comprehension and acquisition) has well been delineated by VanPatten (2012). In his information processing model, he suggests principles of input processing as follows:

(a) *Primacy of Content Words*: Learners process content words in the input before anything else;
(b) *Lexical Preference Principle*: If grammatical forms express a meaning that can also be encoded lexically (e.g., that a grammatical marker is redundant), then learners will not initially process those grammatical forms until they have lexical forms to which they can match them;

(c) *First Noun Principle*: Learners tend to process the first noun or pronoun they encounter in a sentence as the subject;
(d) *Lexical Semantics Principle*: Learners may rely on lexical semantics, where possible, instead of the First Noun Principle to interpret sentences.
(e) *Event Probabilities Principle*: Learners may rely on event probabilities, where possible, instead of the First Noun Principle to interpret sentences (pp. 270–272).

As the principles above indicate, L2 learners tend to rely on lexical-semantic processing and hence need to be helped to reallocate their limited attentional resources strategically to grammatical forms; otherwise, necessary relations between form and meaning/ function for acquisition can hardly be made.

Sharwood Smith (1986) also highlights the importance of simultaneous semantic and syntactic processing as a prerequisite for L2 learning. The double arrows in Figure 3 represent acquisition-related processing, whereas the single arrows indicate communication/ comprehension-related processing. As the figure indicates, L2 competence develops through iterative comparisons and adjustments of the semantic representation and the surface structure of the input. That is, any discrepancies detected between semantic representation and the surface form drive L2 learners to restructure their current IL system by adjusting those two components.

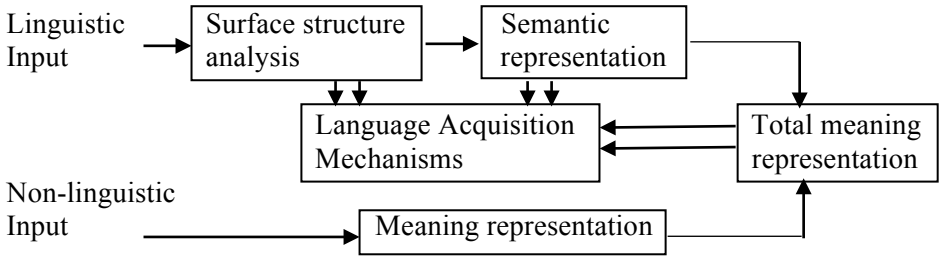

Figure 3. Model of input processing and dual relevance (*Source*: Adopted from Sharwood Smith, 1986)

Also, learners are able to extract meaningful propositions in a top-down manner, relying on non-linguistic cues and their background knowledge. In this case, while comprehension can be achieved, there is only a slim chance of L2 development occurring due to a lack of syntactic processing. On top of that, adult L2 learners have to

deal with their maturational constraints while their L1 system wields a continuous interfering influence, often termed language transfer (Han & Cadireno, 2010; Jarvis & Pavlenko, 2008; Odlin, 1989).

Nevertheless, as is evident in the dominance of communicative language teaching and the increasing popularity of content-based reading instruction, it is clear that comprehension has long been the central construct in L2 reading instruction, whereas how to promote syntactic processing for L2 development during reading has by and large been overlooked (Alderson, 1984; Pica, 2002; Urquhart & Weir, 1998; Zyzik & Polio, 2008). Putting an exclusive emphasis on comprehension reinforces learners' propensity to process input for meaning and thereby may potentially inhibit, if not preclude, the occurrence of the types of syntactic processing necessary for L2 acquisition (Sharwood Smith, 1986; VanPatten, 1996). This problem has been further magnified by the over-adoption of L1 reading research in L2 reading studies, with the distinct complexity of L2 acquisition left largely unattended (Bernhardt, 2005; Han et al., 2009).

Having recognized the limitations of comprehension-exclusive instruction, several eclectic approaches have been proposed to subsume the L2 acquisition-oriented focus into L2 reading instruction, with the primary concern still being comprehension (Leow, 2009; Long, 1991; Sharwood Smith, 1993; VanPatten & Cadierno, 1993). Textual modification techniques such as *simplified input* (Blau, 1982), *textual input enhancement* (Sharwood Smith, 1993) and *glossing* (Johnson, 1982) are but some instantiations of such eclectic approaches. As a review of these approaches will later reveal, controversy remains as to how to strike a balance between developing L2 reading skills and developing L2 competence (Leow, 2009), particularly due to the inconclusive findings for the relationship between L2 reading comprehension and L2 acquisition (Grabe, 2005, 2009; Pulido, 2007, 2009). In other words, as Grabe (2009)

aptly pinpoints, “it is nonetheless difficult to build strong linkages between the development of reading comprehension as a skill and SLA with its strong emphasis on linguistic representation” (p. 204).

2. Role of attention and awareness

The common underlying assumption of the aforementioned text modification approaches to L2 reading is that it is crucial to assist learners to channel their attention towards TL constructions while reading for meaning comprehension. Hence, this section presents a brief overview of the major models on the role of attention and awareness in SLA (Leow, 2015; Robinson, 1995c; Schmidt, 1990, 2001; Tomlin & Villa, 1994), which have served as a theoretical ground for the use of textual modification techniques to promote L2 learning.

2.1. Tomlin and Villa’s functional model of attention

Motivated by Posner (1988; Posner & Petersen, 1990), Tomlin and Villa (1994) suggest a fine-grained analysis of attention that consists of separate yet interrelated neurological networks, whose components include *alertness*, *orientation* and *detection*. Alertness represents an overall readiness to respond to incoming stimuli or data. While alertness does have some relevance to SLA, in that learners generally need to be ready to attend to input, of more importance is orientation, which pertains to directing attentional resources to specific sensory information. Orienting and specific aligning of attention further facilitates detection, i.e., cognitive registration of sensory stimuli. Tomlin and Villa propose that detection is central to language learning, as it is the stage where a particular and specific piece of information is selected, consuming full attentional resources in that moment, and becomes available for further processing, such as hypothesis formation and testing (Posner & Petersen, 1990). Thus, in this model, only detection is necessary for L2 learning, while alertness and orientation may enhance

the possibility of detection occurring. In Tomlin and Villa's account, awareness refers to "a particular state of mind in which an individual has undergone a specific subjective experience of some cognitive content or external stimulus" (p. 193). Drawing on Allport's (1988) criteria to determine awareness, Tomlin and Villa assert that none of the three components of attention – alertness, orientation, and detection – requires awareness, either to function or as a result of processing. In other words, in their view, learning is possible without awareness.

2.2. Schmidt's Noticing Hypothesis

In contrast to Tomlin and Villa (1994), Schmidt (1990, 1995, 2001) argues that awareness is necessary for deeper processing of input in order for learning to take place. In his view, learning entails more than mere subliminal perception of stimuli, as it requires a certain level of awareness. He proposes two levels of awareness: *noticing* and *understanding*. The lower level of awareness is noticing (focal attention) where stimuli are subjectively experienced. In his words, "noticing is the necessary and sufficient condition for converting input to intake" (p. 129), which evolved into the *Noticing Hypothesis*. The higher level of awareness is understanding, where noticed information is analysed and its significance is compared to other information. Hence, Schmidt argues that awareness at the level of understanding is needed for hypothesis testing and the analysis of complex stimuli. While researchers generally support the importance of noticing for learning a single item or simple and reliable rules (e.g., DeKeyser, 2005; Hulstijn, 1995; Leow, 2009), the role of awareness in learning irregular, unreliable and pattern-based features has not been free from debate. Schmidt (2001) also proposed a weaker version of the hypothesis and modified the role of noticing as a facilitator, not as a prerequisite of L2 development.

2.3. Robinson's model of attention and awareness

After reviewing research into the role of attention, memory and their relationship to SLA, Robinson (1995c) supports Schmidt's (1990, 1995, 2001) Noticing Hypothesis, claiming that learning requires some level of awareness. At the same time, his model incorporates Tomlin and Villa's (1994) notion of detection. According to Robinson (1995c), noticing is conceived as "detection plus rehearsal in short-term memory, prior to encoding in long-term memory" (p. 296). More specifically, learning begins with the detection of stimuli accompanied by activation of short-term memory, which is followed by rehearsal to store the stimuli long enough to reach the level of awareness. As a result of rehearsal, a mental trace is left in long-term memory, and input transforms into intake. Thus, in his account, intake is "what is both detected and then further activated following the allocations of attentional resources from a central executive" (Robinson, 1995c, p. 297). Robinson also proposes that learners' linguistic performances are influenced more by the consciously controlled processing demands of a task, rather than by consciously or unconsciously accessible knowledge. Corresponding to the external demands posed by a certain task, different processing mechanisms may operate, which in turn will affect the nature of encoding the stimuli. That is, Robinson highlights the need to consider the attentional demands of a task as modulators of the extent of noticing, which in turn affects SLA.

2.4. Leow's model of the L2 learning process in instructed SLA

Synthesizing various models of attention and awareness in L2 learning (e.g., Chaudron, 1985; Gass, 1997; Leow, 2001a; McLaughlin, 1987; Robinson, 1995c; Schmidt, 1990; Swain, 2005; Tomlin & Villa, 1994; VanPatten, 2004), Leow proposes a model of L2 learning processes premised on the crucial role of attention in SLA. The model includes three major processing stages: an input processing stage, an intake

processing stage, and a knowledge processing stage. These processing stages correspond to stage 1, stage 3 and stage 5 in Figure 4, respectively.

Depending on the level of attention (peripheral, selective or focal), the input processing stage (stage 1) is divided into three sub-phases, namely, attended intake, detected intake and noticed intake. Attended intake is a product of peripheral attention, which is comparable to Chaudron's (1985) concept of the initial stage of perception of input. Attended intake, however, is most likely to be discarded without further storage or processing in the learner's working memory. Selective attention to input, accompanied by a very low level of processing, results in detected intake, which is in line with Tomlin and Villa's (1994) notion of detection. When input is cognitively registered with focal attention, combined with a low level of awareness, it is converted into noticed intake, which is equivalent to Schmidt's (1990) notion of noticing. Noticed intake has the most potential to be incorporated into the learner's developing L2 grammar system.

In the intake-processing stage (stage 3), preliminary intake is subjected to conceptually-driven or data-driven processing. Conceptually-driven processing is accompanied by a higher level of awareness, which enables the conscious encoding or decoding of preliminary intake through activating relevant prior knowledge. Data-driven processing also involves encoding and lodging incoming intake into the developing L2 system, but in a linguistically unsystematic fashion. The developing L2 system, therefore, includes unsystemised linguistic knowledge based on discrete, item-based data as well as systemized linguistic knowledge rooted in internalized or learned data. In addition, a low level of processing of intake may result in implicit restructuring of the learner's developing L2 system, whereas a higher level of awareness enables processing of intake for explicit learning, such as hypothesis testing or rule formation. In other words, depending on the depth of processing, the level of awareness and the

amount of cognitive effort, the same linguistic data can be processed for either implicit or explicit learning.

The knowledge-processing stage (stage 5) takes place between the developing L2 system and the learner's production of output. The appropriateness and solidity of knowledge can be observed in the fluency and accuracy of learners' L2 production. The knowledge-processing stage demonstrates that L2 learning process is not linear, as learners constantly modify their L2 knowledge through monitoring their production, use feedback in order to confirm or disconfirm their L2 knowledge, and utilize their own output as additional input.

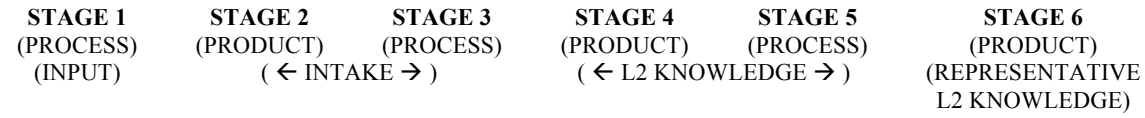

Figure 4. Model of the L2 learning process in instructed SLA (*Source*: Adopted from Leow, 2016, p. 241)

2.5. Shared understanding

The above models posit different views on the importance of awareness in learning. More specifically, Tomlin and Villa propose that awareness does not play a central role in the input-to-intake stage, whereas Schmidt, Robinson, and Leow consider awareness to be an important step for input to be converted into intake. Nevertheless, there is general acceptance among researchers that attention plays a facilitative role in SLA. Without attention, there will be little, if any, learning, as unattended stimuli may remain in short-term memory for only a few seconds and not be available for further processing (Robinson, Mackey, Gass, & Schmidt, 2012; Godfroid & Uggen, 2013; Godfroid, Housen, & Boers, 2010; Godfroid, Boers, & Housen, 2013). Premised on this shared understanding, researchers have investigated whether textual modification techniques can steer learners' attention towards target linguistic constructions while

reading for comprehension. The following section presents a review of empirical research into the efficacy of textual modification.

3. Textual modifications to promote L2 learning from L2 reading

Textual modification has been a common practice among language teachers, textbook writers and curriculum developers. It is not very difficult to find a reading passage for L2 learners accompanied by definitions of unfamiliar words in the margin (i.e., glossing) or with target linguistic items embedded in the passage made salient through typographical manipulation (i.e., textual input enhancement). Also, texts are often shortened and linguistically simplified in order to increase their readability and promote comprehension (i.e., textual simplification), especially when it comes to lower proficiency learners. However, whether these pedagogic strategies do indeed have facilitative effects on L2 learning has not been fully confirmed by empirical findings, and researchers are still debating and examining the efficacy of these methods for promoting development in the use of various L2 linguistic features. This section presents a review of studies that have explored the effects of three common textual modification techniques, i.e., textual simplification, textual enhancement and glossing, on L2 learning from L2 reading.

3.1. Textual simplification

The theoretical premise of textual simplification is that linguistically simplified input may become more readily comprehensible to learners, and thereby indirectly lead them to allocate spare attentional resources to TL form-meaning connections contained in a text (e.g., Hatch, 1983; Long, 1985). This postulation fits neatly into the cognitive theory of limited attentional capacity, in which humans are presumed to suffer from cognitive overload when pushed to simultaneously process multiple information drawing from the same cognitive resource pool. On top of that, simplification is

assumed to assist learners' transition from the initial stage of learning, where consciously controlled and capacity-consuming text processing taxes their limited attentional resources, to a more advanced stage of learning where text processing is automatized through repeated exposure and practice, hence requiring less mental effort (McLaughlin, 1987).

According to Leow (1993), simplified input can be defined as "second language input that has been modified by a speaker/ writer to facilitate second language learners' comprehension … [and] include phonological (on oral input), morphological, syntactic, lexical, and discourse modification" (p. 334). The methods used for simplification involve using high frequency vocabulary, fewer idioms, fewer pronouns, simpler syntactic structures and reducing the length of a text, to name but a few (Leow, 1997b, 2009). Additionally, length, topic and genre of texts have differed among these studies (e.g., Blau, 1982; Davies, 1984; Doddis, 1985; Leow, 1993; Oh, 2001; Wong, 2003; Yano, Long, & Ross, 1994; Young, 1999). The effects of text simplification on comprehension have been measured via diverse types of tasks, such as multiple-choice questions (e.g., Blau, 1982; Oh, 2001; Yano et al., 1994; Young, 1999), cloze tests (e.g., Davies, 1984) and written free recall (e.g., Wong, 2003; Young, 1999). While some studies found no significant effects of simplification on comprehension (Doddis, 1985; Young, 1999), others reported facilitative effects for simplified input on comprehension (Blau, 1982; Davies, 1984; Oh, 2001; Wong, 2003; Yano et al., 1994). As Leow (2009) aptly summarizes, methodological differences between studies, such as the number and characteristics of texts used, the duration of treatment, types of assessment tasks and learner's L2 proficiency, might have contributed to the inconclusive findings. What is more, "the open-ended approach to what constitutes simplification" (Leow, 1997c, p. 294) has yielded inconsistent findings.

There are also studies that have compared the effects of simplification with those of elaboration (Oh, 2001; Yano et al., 1994). Elaboration can be defined as modifying a text for easier comprehension by adding redundant information by means of repetition, paraphrases and appositionals (Long, 1996). Yano et al. (1994) tested whether elaborated input with increased levels of redundancy and explicitness to compensate for unknown linguistic items could serve as a potential alternative to text simplification. In this study, participants were presented with one of three types of texts: native baseline, simplified and elaborated. The scores on a comprehension test, measured by 30 multiple-choice test items, were not significantly different between the simplified and the elaborated versions. It was also found that different types of text modification interacted with different levels of comprehension, i.e., *replication* that requires readers to build a mental reproduction of the textual information with no or minor changes; *synthesis* that requires readers to connect multiple ideas that are explicitly expressed across different parts of a text; and *inference* that requires readers to make a deduction about the implications of a text (Long & Ross, 1997; Yano et al., 1994). More specifically, scores on replication items were highest among the readers of simplified texts, whereas those for inference were highest among the readers of elaborated texts. Based on these findings, Yano et al. (1994) suggest elaboration should be favoured over simplification, as simplification involves the removal of linguistic qualities that learners eventually need to learn. Oh's (2001) study, which adopted a design comparable to Yano et al.'s study, produced similar results, implying that text modification may need to be in the direction of elaboration, as native-like features can be maintained while being equally as successful as simplification in facilitating comprehension. However, it remains subject to debate whether equivalent comprehensibility corresponds to equivalent learning opportunities because, as discussed earlier, comprehension is by no

means isomorphic to acquisition (Sharwood Smith, 1986, 1991, 1993; VanPatten, 1990, 1996, 2004).

Of more relevance to this current thesis are studies that investigated the effects of text simplification on learners' intake from texts (Leow, 1993; Wong, 2003). Leow's (1993) study was the first to explore whether simplification facilitated learners' noticing of TL form-meaning mapping. The participants were 137 university students learning Spanish. The target forms were the Spanish present perfect and present subjunctive. Two passages containing equal numbers of the target forms were simplified, following Hatch's (1983) guidelines for simplification. To ensure different degrees of comprehensibility of the two versions, a pilot study was conducted with another group of university students, and the results showed that the simplified versions were significantly more comprehensible than the unsimplified versions. Intake was operationalised as a correct response to a timed multiple-choice recognition task immediately following a reading task. Data analysis revealed that the participants who read simplified texts did not show significantly more intake than those who read unsimplified texts. Based on these findings, Leow suggests that comprehensibility may not be directly related to acquisition and that external textual manipulation may not be haphazard but inadequate to promote learners' intake. It should be noted, however, that his study involved only a one-time exposure to the texts and no delayed posttest was administered.

Wong (2003) also investigated the effects of simplified input on the acquisition of L2 grammar. More specifically, this study examined whether textual enhancement and simplification affected adult learners' acquisition of French past participle agreement in relative clauses and comprehension of three reading texts. Eighty-one participants were randomly assigned to one of four conditions: (a) textual enhancement and simplification, (b) textual enhancement only, (c) simplification only, and (d) no textual modification.

The study included three treatment sessions, each of which consisted of reading for ten minutes and producing a free written recall of the text. An error correction task was used as a pretest and posttest. The participants also completed a post hoc questionnaire asking if they had noticed any typographical cues in the texts. The results revealed that neither textual enhancement nor simplification had significant effects on the participants' performance on the error correction task. Also, textual enhancement did not affect comprehension, while simplification was shown to have significantly positive effects on comprehension. Based on these findings, Wong concluded that more comprehensible input through simplification might not be sufficient to facilitate acquisition of the target form-meaning connections.

3.2. Textual enhancement

Textual enhancement attempts to steer learners' attention towards target linguistic features during reading by making the features perceptually salient through typographical manipulation, such as underlining, colouring, bold facing, italicizing and capitalizing (Sharwood Smith, 1993). As the premise underlying textual enhancement is that learners must first comprehend what they read before their attention is directed towards form-meaning connections, textual enhancement purports to engage in implicit and unobtrusive intervention to minimize any interruption of comprehension caused by typographical changes. Yet, whether textual enhancement does indeed promote learners' attention to target form-meaning mapping without deleterious effects on comprehension, thereby precipitating ultimate learning, is an empirical question, and research thus far has yielded inconclusive results. While some studies have revealed significant positive effects for textual enhancement (e.g., Doughty, 1991; Jourdenais, Ota, Stauffer, Boyson, & Doughty, 1995; Shook, 1994; Williams, 1999), some have shown only partial effects (e.g., Alanen, 1995; Izumi, 2002; Park, 2004; J. White, 1998) or no significant effects at all (e.g., Bowles, 2003; Lee, 2007; Leow, 1997c, 2001b;

Leow, Egi, Nuevo, & Tsai, 2003; Overstreet, 1998; Wong, 2003). Considering the studies summarized in Table 2, it takes only a little reflection to notice the methodological idiosyncrasies among the studies that could have contributed to inconclusive findings. More specifically, studies differ in terms of the nature of target features (e.g., in their semantic, syntactic and functional complexity, and perceptual saliency), the duration of treatment, measurement methods employed, number of participants, participants' developmental readiness (with or without prior knowledge of the target form-meaning mapping), the frequency of enhanced target items per text and the length of texts (Han, Park, & Combs, 2008).

As aforementioned, textual enhancement presupposes that (a) increasing perceptual salience will drive learners' attention towards enhanced form-meaning connections, and that (b) learning of the attended form-meaning connections will occur provided that attention is what converts input into intake (Izumi, 2002). Even so, much research into textual enhancement does not include independent measures of learners' attention or noticing while being exposed to enhancement, which reduces the internal validity (Leow et al., 2003). Han et al. (2008) also pinpoint that most researchers seem to equate the efficacy of textual enhancement with observable performance in a posttest, while the effects of textual enhancement on noticing have not been properly operationalised and independently researched. In other words, only a few studies have attempted to measure whether and how learners process enhanced input (Alanen 1995; Bowles 2003; Izumi 2002; Jourdenais et al. 1995; Leow 2001b, Leow et al., 2003; Park 2004).

For example, Izumi's (2002) study indicates that heightened noticing induced by textual enhancement might not correlate with more learning. This study set out to investigate the relative effectiveness of output production and textual enhancement during reading for comprehension. The target structure was English relativization, and

61 participants were randomly assigned to one of the following three conditions: output production, textual enhancement and control. During reading, the participants were instructed to take notes on the part of the text they thought important, which served as an indicator of their noticing. A sentence-combination test, a picture-cued sentence completion test, an interpretation test and a grammatical judgment test (GJT) were employed to measure the amount of learning, if any. The results from analyses of the notes taken and posttest scores revealed that output production served as a priming device, proving to be more effective than textual enhancement or control conditions. More importantly, it was also shown that while textual enhancement resulted in significantly more noticing (i.e., note-taking) than the control condition, it was not borne out in posttest scores. Based on these findings, Izumi concluded that noticing might not necessarily result in learning. It is debatable, though, whether note-taking can serve as a sensitive measure of noticing: for one, what learners do not report in their notes cannot be measured, even when noticing occurs; in addition, note-taking can serve as an output-production activity, affecting what learners notice and thereby having a confounding effect on the results.

Similar findings were obtained from Winke's (2013) recent replication of Lee's (2007) study on the effects of textual enhancement and topic familiarity on learners' reading comprehension and learning of English passive constructions. In this study, eye-tracking technology was employed to measure the participants' attention in a more sensitive and accurate way, based on the assumption that learners' mental effort may materialize in the form of longer eye fixation on a text. The results demonstrated that learners looked at the enhanced forms for longer and revisited them more often, supporting textual enhancement as an effective prompt to trigger learners' noticing of target form-meaning connections. Yet, increased noticing was not followed by gains in an immediate test assessing knowledge of English passive constructions. Based on this

finding, Winke suggested that an "increase in the amount of noticing may not be enough for immediately measurable acquisition to occur" (p. 341). This finding seems to be in line with one argument that "forms may be noticed perceptually, but not linguistically" (Leeman, Areagoitia, Fridman, & Doughty, 1995, p. 219). Godfroid, Housen, and Boers (2010), based on their findings from eye-movement data, also claimed, "*it is the type of mental elaboration that follows noticing* [original emphasis] which determines the strength of the memory trace created and, thus, determines eventual learning gains" (p. 186). Winke further asserted that a lack of clear goals for reading could have failed to provide learners with true motivation to read, which resulted in a lack of learning induced by textual enhancement. She also suggests that future studies may need to investigate the covarying effects of task goals (e.g., reading directions, task objectives, preemptive vs reactive instructions) and textual enhancement on noticing and learning from reading. This coincides with other researchers' suggestion that task requirements may affect what is attended to and noticed during learners' on-line processing, thus moderating the effects of input modification (Doughty, 1991, 2001; Doughty & Williams, 1998; Robinson, 1995b; Skehan, 1996, 1998). In other words, by pushing learners further by providing them with explicit task goals, noticed target form-meaning connections via enhancement might be processed at a deeper level, resulting in greater gains.

By contrast, Leow (2001b), Leow et al. (2003) and Bowles (2003), who operationalised noticing as incidents of reporting the target form-meaning mapping in concurrent think-aloud protocols, do not support the assumption that textual enhancement facilitates noticing. Most notably, in Leow's (2001b) study, 74 university students read a text wherein Spanish imperatives were textually enhanced, while verbalizing their thinking. Comprehension was measured with short-answer and multiple-choice comprehension questions, and intake was operationalised as scores

from a multiple-choice test and a fill-in-the-blanks test. The results of this study revealed that there was no significant difference in the amount of noticing between the textual enhancement and control groups, while the amount of noticing was shown to correlate with intake in both the enhanced and unenhanced groups. Based on these findings, Leow cast doubt on the efficacy of textual enhancement to promote noticing and increase the amount of intake.

In a similar vein, in a recent eye-tracking study conducted by Indrarathne and Kormos (2016), textual enhancement played only a marginal role in noticing of and development in the knowledge of the English causative *had*. One hundred participants were divided into a control group and four experimental groups: input flood, textual enhancement, a specific instruction to pay attention to the target construction, and an explicit metalinguistic explanation of the target construction. Eye-movement data were obtained from 45 participants in the sample. Textual enhancement was achieved through boldfacing all instances of the target feature. To measure processing of the target construction, two eye-movement measurements (i.e., total fixation duration on each occurrence of the target construction and mean total fixation duration for all occurrences) were taken for each participant. Knowledge of the causative *had* was assessed with a sentence reconstruction task and a grammaticality judgment task. The results indicated that eye-movement indices increased significantly for the specific instruction to pay attention to the target construction and the metalinguistic instruction groups. The participants in these groups also exhibited significantly improved knowledge of the target structure. With respect to the limited efficacy of textual enhancement on eye-movements and posttest scores, Indrarathne and Kormos assumed that boldfacing might not have been as effective as Winke's (2013) textual enhancement through underlining. In sum, research findings have yielded an inconclusive picture as to the effects of textual enhancement on reading comprehension and noticing and/or

learning of target L2 constructions, mainly due to the methodological idiosyncrasies in existing research as well as lack of conceptual and operational rigour and consistency across studies, thus highlighting the need for more empirical studies.

Table 2. Summary of the studies on textual enhancement

Study *N* = Number of participants	**Target form(s)**	**Duration of the treatment**	**Measurement** P: Process, L: Learning, C: Comprehension	**Results**
Doughty (1991) *N* = 20	English relative clauses	Ten sessions	- C: Comprehension questions, free recall task - L: GJT; sentence combination task, guided sentence completion task; oral task	Positive effects on acquisition
Shook (1994) *N* = 125	Spanish present perfect/ relative pronouns	Two sessions, less than 1 hr each	- L: Recognition task; fill-in-the-blanks production task	Positive effects on acquisition
Alanen (1995) *N* = 36	Finnish locative suffixes/ consonant gradation	Two sessions, 15 minutes each	- L: Sentence completion task; GJT; rule statements	Positive effects on acquisition
Jourdenais et al. (1995) *N* = 10	Spanish preterit/imperfect	One session, less than 1 hr	- P: Think-aloud protocols - L: Picture-based writing task	Positive effects on noticing/ intake
Leow (1997b) *N* = 84	Spanish formal imperatives	One session, less than 1 hr	- C: Short-answer comprehension task - L: Multiple-choice form recognition	No effects on acquisition
J. White (1998) *N* = 86	English possessive determiners	Six sessions, 10 hrs in total	- L: Passage correction task; multiple-choice test; Oral picture description task	Partial effects on acquisition
Overstreet (1998) *N* = 50	Spanish preterit/ imperfect	One session, less than 1 hr	- C: T/F comprehension quiz - L: Circle-the-verb task; written narration task	No effects on acquisition
Williams (1999) *N* = 58	Italian possessive adjectives/ inflectional verb endings for subjects	One session, about 2 hrs	- L: Verbatim memory task; translation task	Positive effects on acquisition
Leow (2001b) *N* = 38	Spanish imperatives	One session, less than 1 hr	- P: Think-aloud protocols - C: Short-answer and multiple-choice tasks - L: Multiple-choice recognition task; fill-in-the-blanks production task	No effects on noticing/ intake
Izumi (2002) *N* = 61	English relativization	Six sessions, 30–60 minutes each one	- P: Note-taking - L: Sentence-combination test; picture-cued sentence completion test; interpretation test; GJT	Positive effects on noticing (note-taking), but no effects on acquisition
Bowles (2003) *N* = 15	Spanish imperatives	One session, less than 1 hr	- P: Think-aloud protocols - C: Short-answer and multiple-choice tasks	No effects on noticing or intake

			- L: Multiple-choice recognition task; fill-in-the-blanks production task	
Leow et al. (2003) *N* = 72	Spanish present perfect	One session, less than 1 hr	- P: Think-aloud protocols - C: Multiple-choice comprehension task - L: Multiple-choice recognition task	No effects on noticing/ intake
Wong (2003) *N* = 81	French past participle agreement in relative clauses	Three sessions, less than 1 hr each	- C: Free recall task - L: Error correction task	No effects on acquisition
Park (2004) *N* = 24	English reporting past events (verb-backshifting)	Two sessions, 35 minutes each one	- P: Think-aloud protocols - L: Written picture description task	Partial effects on noticing/ acquisition
Kim (2006) *N* = 297	English vocabulary	One session, 20 minutes	- L: Form-recognition and meaning-recognition vocabulary test	Partial effects on meaning recognition
Lee (2007) *N* = 295	English passive	Four sessions, 50 minutes each one	- C: Free recall task - L: Form correction tasks	Positive effects on acquisition, but unfavourable effects on comprehension
Winke (2013) *N* = 55	English passive	One session Less than 1 hr	- P: Eye-tracking - C: Free recall task - L: Form correction tasks	Positive effects on noticing, but no effects on acquisition
Park & Nassif (2014) *N* = 16	Arabic comparative, dual pronoun	Two sessions, 10 minutes each one	- C: Free recall task - L: Fill-in-the blank task, free production task	No effects on acquisition, negative effects on comprehension
Jahan & Kormos (2015) *N* = 97	Will, be going to	Two sessions Less than 1 hr	- P: Noticing questions - C: Multiple-choice test - L: Metalinguistic task, fill-in-the-blank task, multiple-choice task	Positive effects on noticing, no effects on acquisition and comprehension
LaBrozzi (2016) *N* = 109	Spanish preterit tense of *-er*	One session, 10 minutes	- C: Multiple-choice test - L: Translation task	Positive effects on form recognition, no effects on comprehension
Loewen & Inceoglu (2016) *N* = 30	Spanish preterit, Imperfect	One session, 15 minutes	- P: Eye-tracking - L: Form production task, oral picture description task	No effects on noticing/ acquisition
Indrarathne & Kormos (2016) *N* = 100	Causative *had*	Three sessions Less than 1 hr	- P: Eye-tracking - C: Multiple-choice test - L: Sentence reconstruction task, grammaticality judgment task	No effects on noticing

3.3. Glossing

A gloss can be defined as information provided about an unfamiliar linguistic item, in the form of a definition, synonym or translation, to reduce linguistic obscurity and thereby promote better comprehension. As the following review will demonstrate, various glossing techniques have been used, such as L1 versus L2 glosses (e.g., Ko,

2012), computerized versus paper-based glosses (e.g., Bowles, 2004), textual versus multimedia glosses (e.g., Chun & Plass, 1996) and single versus multiple-choice glosses (e.g., Rott, 2005), among others. Moreover, the degree of metalinguistic explicitness in the glosses provided also differs across studies, such as a simple definition or a synonym of a word (e.g., Guidi, 2009), a combination of definition and its use in an exemplar sentence (e.g., Hulstijn & Laufer, 2001), a combination of a translation and a relevant picture or video clip (e.g., Al-Seghayer, 2001), to name but a few. Probably due to the divergent operationalisation of glossing, the results from research into the role of glosses in L2 reading comprehension and L2 vocabulary learning have been inconclusive (see Table 3).

3.3.1. Glossing and L2 reading comprehension When it comes to the efficacy of glossing in L2 reading comprehension, some studies do not support a facilitative role for it (e.g., Bell & LeBlanc, 2000; Davis & Lyman-Hager, 1997; Jacobs et al., 1994; Johnson, 1982; Lomicka, 1998; Pak, 1986), whereas others have found positive results for L2 reading comprehension (e.g., Bowles, 2004; Chun & Plass, 1996; Ko, 2005; Martinez-Fernández, 2010). Aside from the aforementioned types of glossing, various comprehension measurements might have contributed to the divergence in previous findings. Measures previously used include cloze tasks (e.g., Pak, 1986), free recall tasks in L1 (e.g., Bell & LeBlanc, 2000), free recall tasks in L2 (e.g., Johnson, 1982), multiple-choice comprehension items (e.g., Ko, 2005), comprehension questionnaires (e.g., Martinez-Fernández, 2010) and think-aloud protocols (e.g., Lomicka, 1998).

As reviewed in an earlier part of this thesis, reading comprehension generally involves conceptual representations with several mutually constraining layers, typically a local-level representation based on text-based information and a global-level representation relying on textual information and the reader's background knowledge

(Khalifa & Weir, 2009; Kintsch, 1998). The level at which glossing is purported to function is local, through easing bottom-up meaning extraction and assisting in the construction of semantic propositions. If glossing does fulfil this function, the reader will not only build a more accurate text-model, but also establish a richer situation model due to the mental resources saved from trying to infer unknown words. That said, it seems important to measure both local comprehension (i.e., comprehension of the phrase, clause or sentence containing the glossed item) as well as global comprehension, and to examine how glossing makes a distinctive contribution to each.

Several researchers have highlighted this issue (e.g., Johnson, 1982; Ko, 2005; Lomicka, 1998). For example, Ko (2005) suggested that glossing might promote learners' use of inferring strategies, resulting in better understanding of a text. This study consisted of a quantitative analysis of scores from multiple-choice comprehension questions collected from 94 Korean learners of English and a qualitative analysis of additional 12 participants' think-aloud protocols. The participants were randomly assigned to one of three conditions: L1 gloss, L2 gloss and control. The results of quantitative analysis revealed that scores from the L2-gloss group were significantly higher than those of the control group. Also, qualitative analysis demonstrated that both gloss groups reported more higher-level strategies, whereas the control group seemed to be predominantly bound up in bottom-up analysis. In other words, glossing seemed to help the participants free up their limited mental resources for inference processing, whereas those in the control group were struggling with decoding. In contrast, in Johnson's (1982) study, marginal glosses were found to have no impact on comprehension. Johnson suggested that glossing might have led L2 readers to focus more on bottom-up processes by encouraging word-by-word decoding, and thus have a detrimental effect on text comprehension. Hence, when investigating the role of glossing in L2 reading, it seems worthwhile to decompose the construct of

comprehension into multiple levels and investigate whether glossing has a facilitative or deleterious impact on each level of local and global comprehension.

3.3.2. Glossing and L2 vocabulary acquisition Recent research on glossing has also directed its focus to the effectiveness of glossing on promoting the incidental acquisition of L2 vocabulary, based on the assumption that glossed lexical items may receive heightened attention from the learner without interrupting the reading process, resulting in incidental acquisition of the form-meaning mapping of the lexical items. Previous studies have shown that, overall, glossed texts have positive, though small, effects on L2 vocabulary learning (e.g., Bowles, 2004; Hulstijn, 1992; Hulstijn, Hollander, & Greidanus, 1996; Nagata, 1999; Watanabe, 1997), while the relative efficacy of different types of glosses may vary (e.g., Bowles, 2004; Chun & Plass, 1996; Guidi, 2009; Hulstijn & Laufer, 2001; Kost, Foss, & Lenzini, 1999; Martinez-Fernández, 2010; Nagata, 1999; Rott, 2005). For example, Bowles (2004) compared the respective efficacy of paper-based versus computer-based L1-glosses on the noticing of glossed items, the comprehension of text, and subsequent recognition and production of words. The participants were 50 university students learning Spanish. Comprehension was measured via multiple-choice comprehension questions in the L1, and intake was operationalised as the correct selection of the meaning of a target word out of four options in a multiple-choice recognition posttest. A written translation posttest was used to measure the participants' ability to produce the targeted words. Both recognition and production posttests were administered immediately after, and three weeks after, the treatment. The results of the posttest scores revealed significant roles for both types of glossing on comprehension and intake, whereas no difference was found between the two conditions (paper-based vs computer-based). Also, the analysis of think-aloud protocols demonstrated that the target vocabulary items were noticed significantly more in glossed conditions than in the control condition. Based on these findings, Bowles

suggested that glossed texts, regardless of mode, aided the understanding of text as well as drawing learners' attention to targeted words. The think-aloud protocols also revealed a low level of awareness of the form-meaning connections of target items in general, which appears to be in line with previous studies showing a marginal influence of text modifications in promoting the acquisition of target L2 features (e.g., Leow, 2001b, Leow et al., 2003).

With respect to the overall small gains in L2 vocabulary learning from glossing, some researchers argue that providing readily accessible word meanings may have an insignificant, or even prohibitive, influence on L2 vocabulary learning, as L2 readers are not encouraged to invest any mental effort to get the meanings of words (Hulstijn, 1992; Hulstijn & Laufer, 2001; Hulstijn, Hollander, & Greidanus, 1996; Laufer & Hulstijn, 2001). Drawing on the notions of depth of processing (Craik & Lockhart, 1972) and the degree of elaboration (Craik & Tulving, 1975), Laufer and Hulstijn proposed the *Involvement Load Theory* whose components entail a *need* to complete a reading task, a *search* to infer the meaning of an unknown word, and an *evaluation* of the semantic and grammatical fit of an inferred meaning in context. Depending on the presence or absence as well as the intensity of these three components, learners may engage in differing levels of involvement load, which in turn affects the level of processing and the robustness of retention of the vocabulary item. However, researchers who support glossing reject this argument as implying that ability varies greatly among learners, and it is also possible that the inferences made by learners are not always reliable due to deceptive or insufficient contextual information (e.g., Beck, McKeown, & McCaslin, 1983; Gettys, Imhof, & Kautz, 2001; Watanabe, 1997). As such, whether forcing learners to engage in inference/ search processes (in Hulstijn and Laufer's words, greater involvement load) has a stronger impact than glossing on L2 vocabulary learning has been an empirical question.

Hulstijn and Laufer (2001), in order to test Involvement Load Theory, compared the efficacy of L1 glosses, fill-in tasks and composition tasks on L2 vocabulary retention. In this study, 87 university students in the Netherlands and 99 in Israel were randomly assigned to one of three groups: L1 gloss, fill-in task and composition. In the L1 gloss group, L1 translations of ten target words were provided in the margin of a text, whereas in the fill-in task group, the participants were asked to fill in ten gaps with missing words from a list. Also, in the composition group, the participants were instructed to write a composition using the ten target words, for which explanations and examples of their usage were given. All groups read the same text while answering comprehension questions, and they did an unannounced vocabulary translation task immediately after, and one to two weeks after, the reading task. The results revealed that the composition group significantly outperformed the other groups, while the fill-in group demonstrated significantly higher retention scores than the L1 gloss group. Based on these findings, Hulstijn and Laufer concluded that Involvement Load Theory was empirically supported, manifesting a stronger impact of deeper and more elaborate processing induced from fill-in or composition tasks compared to providing readily accessible word meanings via glosses.

Hulstijn and Laufer's (2001) proposal was supported by the results of further studies motivated by Involvement Load Theory (e.g., Keating, 2008; Kim, 2008; Rott, 2005; Rott & Williams, 2003; Rott, Williams, & Cameron, 2002). For example, in Kim's study where two different proficiency levels were included, the results showed that task effects (i.e., glosses vs fill-in task vs composition task) overrode the influence of proficiency and that composition groups outperformed both fill-in and gloss groups in an immediate as well as a delayed posttest, while the fill-in group outperformed the gloss group in a delayed posttest. Yet, it should be noted that, in these studies, (a) a control group was often not included (except for Keating's study), (b) experimental

conditions differed not only in terms of involvement load, but also the degree of output orientation (composition tasks), and (c) concurrent process measures such as think-aloud were not employed, which could have cast light on the amount of involvement load induced in each task condition.

3.3.3. Glossing and learning of L2 grammatical constructions In contrast to the considerable number of studies on the usefulness of glossing in L2 vocabulary learning, research into the role of glossing in the learning of L2 constructions other than lexis is only scant (e.g., Guidi, 2009; Martinez-Fernández, 2010; Nagata, 1999). In Guidi's (2009) study, 65 learners of Spanish were randomly assigned to one of four conditions: L1 gloss with think-alouds, L1 gloss without think-alouds, no gloss with think-alouds, and no gloss without think-alouds. Immediately after, and three weeks after, finishing a reading comprehension task, participants completed production and recognition tests to measure their knowledge of target constructions, which were ten Spanish lexical items and Spanish present perfect and impersonal *se*. The results demonstrated that glossing promoted reading comprehension, but had only limited influence on development in the knowledge of the target constructions. Guidi suggested that the inherent characteristics of the target features, such as salience, frequency, abstractness of meaning-referent or complexity of encoding, might have interacted with the effects of glossing. Similar results were obtained in Martinez-Fernández's (2010) study that compared the efficacy of glossing and fill-in tasks on the learning of ten Spanish lexical items and Spanish subjunctive constructions. In this study, the participants in the fill-in task condition were not required to write down a missing word but rather to underline one word from two options for each gap. In order to measure the learning of target L2 words and grammatical items, recognition and production posttests were administered immediately after, and one week after, the treatment. The results revealed that both L1 gloss and fill-in task significantly promoted learning of the target

words, but not the Spanish subjunctive construction. In contrast, in Nagata's (1999) study with 26 learners of Japanese, it was found that development in three Japanese grammatical structures (i.e., '*hado*' meaning *as much as*, '*kara*' meaning *because*, and a noun phrase followed by '*to*' meaning *with*) was significant in both single- and multiple-choice glossed conditions. Given the small number of studies on the usefulness of glossing in promoting the acquisition of L2 grammatical constructions, in addition to the methodological differences among those, the accumulation of more empirical findings seems imperative.

What seems noteworthy is that, except for a few studies (e.g., Bowles, 2004; Guidi, 2009; Martinez-Fernández, 2010), learners' cognitive processes while performing reading tasks were often not investigated in spite of their usefulness in documenting learner-internal processes induced by various types of glossing. As in the textual enhancement literature, the lack of independent measures of learners' cognitive processes resulted in assumptions that were based on speculation rather than empirical evidence about the impact of glossing on L2 comprehension and development. For example, some researchers argue that learning of target L2 vocabulary occurred as the learners paid conscious attention to items during reading (e.g., Hulstijn, 1992; Hulstijn & Laufer, 2001; Watanabe, 1997), but this explanation was not confirmed directly with empirical evidence. Considering that the thrust of utilizing glossing for L2 development is premised on the essential role of attention as a mediator of input and intake, it seems necessary to employ independent measures of learners' online reading/ learning processes when investigating their impact on L2 developmental processes.

Table 3. Summary of the studies on glossing

Study *N* = Number of participants	**Glossing type(s)**	**Measurement** P: Process, L: Learning, C: Comprehension	**Results**
Johnson (1982) *N* = 72	Marginal definition	- C: Written recall task; cloze test; multiple-choice questionnaires	No effects on comprehension
Pak (1986) *N* = 65	Marginal definition	- C: Cloze test	No effects on comprehension

Davis (1989) *N* = 71	Marginal, online, annotation glossing	- C: Written recall task	Positive effects on comprehension (against control)
Kost et al. (1991) *N* = 56	*20 target words* L1-textual vs pictorial vs textual+pictorial	- L: Production task; picture recognition task; word recognition task	Significantly stronger effects for combined gloss than textual or pictorial only
Jacobs et al. (1994) *N* = 85	L1 vs L2 definitions	- C: Written recall task	Interaction between glossing and L2 proficiency
Jacobs (1994) *N* = 116	L1 translation	- C: Written recall task	Positive effects on comprehension
Chun & Plass (1996) *N* = 103	*82 target words* Textual pictorial vs video vs control	- C: Comprehension questions - L: Recognition and production posttest	Stronger effects for textual combined with pictorial than textual or video only.
Hulstijn et al. (1996) *N* = 78	*16 target words* L1 gloss vs dictionary use; Frequency of exposure	- C: Comprehension questions - L: Recognition and production posttest	Frequency effects when combined with marginal glosses or dictionary use; greater effects of gloss than dictionary use on recognition test
Davis & Lyman-Hager (1997) *N* = 42	Computerized word definitions	- C: Written recall task; multiple-choice questions	No effects on comprehension
Watanabe (1997) *N* = 231	*16 target words* L2 single gloss vs appositives vs MC gloss vs control	- C: Cloze test; open-ended comprehension questions - L: Production posttest; delayed posttest	L2 gloss and MC gloss outperformed appositives and control group on posttests; no significant difference between L2 gloss and MC gloss
Lomicka (1998) *N* = 12	Computerized word definitions	- P: Think-aloud protocols	No effects on comprehension
Nagata (1999) *N* = 26	*20 target words & 3 grammatical features* L1 gloss vs MC+feedback; frequency (4 levels)	- L: Translation task	MC gloss significantly outperformed L1 gloss
Bell & LeBlanc (2000) *N* = 40	Marginal definition	- C: Written recall task	No effects on comprehension
Al-Seghayer (2001) *N* = 30	*10 target words* Textual vs pictorial vs video (within-subject design)	- L: Recognition and production posttest	Stronger effects for text+video than text+picture, or text only
Gettys et al. (2001) *N* = 22	*43 target words* L1 gloss vs dictionary use	- C: Written recall task; multiple-choice questionnaire - L: Recognition posttest and time on task	Time effects (dictionary-use group spent significantly more time than L1-gloss group)
Hulstijn & Laufer (2001) *N* = 186	*10 target words* L1 gloss vs fill-in task vs writing	- L: L1 translation task; L2 explanation task	Stronger effects for writing than L1 gloss or fill-in gloss
Bowles (2004) *N* = 50	*40 target words* Computer vs paper glossing	- P: Think-aloud protocols - C: Multiple-choice questionnaire - L: Translation production and multiple-choice recognition posttest	Positive effects on comprehension (against control); better performance on recognition than production test

Ko (2005) *N* = 94	L1 vs L2 glosses	- P: Think-aloud task - C: Multiple-choice comprehension questions	Significantly stronger effects for L2 gloss than L1 gloss on comprehension
Rott (2005) *N* = 10	*7 target words* L1 single gloss vs L1-MC gloss	- C: L1 written recall task - L: Word-form recognition test; multiple-choice word meaning recognition test	Stronger effects for MC gloss on recalling supporting ideas and learning target words
Kim (2008) *N* = 64	*10 targeted words,* Graphic organizer vs fill-in task vs composition task	- C: Comprehension questions - L: VKS	Significantly greater gain in the composition task group, no significant difference between graphic organizer and fill-in task groups
Keating (2008) *N* = 79	*8 nonsense words (5 concrete nouns and 3 verbs)* Fill-in task vs composition task vs control	- C: Comprehension questionnaire - L: Passive recall test; active recall test	Retention was highest in the composition task group, lower in the fill-in task group, lowest in the control group
Guidi (2009) *N* = 65	*10 targeted words, Spanish present perfect and impersonal SE* L1-gloss vs control	- P: Think-aloud protocols - C: Multiple-choice comprehension questions - L: Fill-in-the-blanks production test; multiple-choice recognition test	Limited effects; interaction between gloss and type of linguistic item
Martinez-Fernández (2010) *N* = 73	*5 concrete and 5 abstract words, and Spanish subjunctive* L1 gloss vs fill-in task vs control	- P: Think-aloud protocols - C: Comprehension questionnaire - L: Word meaning and grammar production and recognition tests	Positive effects for glossing on vocabulary noticing and learning and text comprehension, no effects on noticing and learning of grammatical items
Ko (2012) *N* = 90	*16 target words* L1 gloss vs L2 gloss vs control	- L: Multiple-choice vocabulary recognition test	Positive effects for glossing on vocabulary learning, no significant difference between L1 and L2 glossing
Huang & Lin (2014) *N* = 118	*8 target words* Inference-gloss-gloss vs gloss-retrieval-gloss vs full glossing	- C: Multiple-choice test - L: Vocabulary form-recall test, meaning-recall test, and meaning-recognition test	Gloss-retrieval-gloss was significantly more effective than other conditions in improving target-word learning

4. Summary

To summarize, L2 reading is a crucial component of SLA, as both an indispensable language skill and a major source of L2 input. How to achieve dual goals, i.e., improving reading comprehension ability and developing L2 competence, has been a prime concern among researchers. So far, various types of text modification techniques, i.e., text simplification, textual enhancement and glossing, have been suggested as potential tools to steer learners' attention towards L2 form-meaning connections during reading for comprehension. The theoretical underpinning of these pedagogical techniques is that attention or noticing is necessary for converting input

into intake, provided that the meaning of the input is comprehended. Yet, previous studies have often not measured learners' reading processes and comprehension outcomes, even though both components are crucial to support internal validity. Thus, it has been suggested that (a) process measures during reading, such as verbal reports or eye-tracking technology, should be used to validate whether textual modification fulfils its purported function, i.e., directing and heightening learners' attention to target L2 features, and (b) both local and global reading comprehension should be measured (Khalifa & Weir, 2009), through replication, synthesis, and inference (Yano et al., 1994), text-model and situation-model (Kinstch, 1998), among others.

Research into textual modification has engendered a muddled picture. The findings from previous studies have revealed that textual simplification has, overall, facilitative effects on L2 reading comprehension (e.g., Blau, 1982; Davies, 1984; Leow, 1993; Oh, 2001; Wong, 2003; Yano et al., 1994), but not on L2 learning (e.g., Leow, 1993; Wong, 2003). Also, research into the role of textual enhancement in terms of facilitating development in L2 competence has exhibited inconsistent findings, while some studies have demonstrated that textual enhancement might have a detrimental influence on comprehension (e.g., Lee, 2007; Overstreet, 1998). Also, glossing has proved to have a positive, though small, impact on reading comprehension and the learning of glossed vocabulary items. While textual enhancement has typically focused on learning L2 grammatical features, glossing has been predominantly researched in association with L2 lexical learning. With respect to the limited efficacy of textual modification, some researchers have proposed incorporating various types of tasks, which may lead learners to read a given text with clearer and stronger motivation (e.g., Hulstijn, 1992, Hulstijn et al., 1996; Hulstijn & Laufer, 2001; Winke, 2013). That is, task goals, instructions and demands might encourage learners to process a given text to a deeper level, thus boosting the likelihood of benefiting from textual modification. One

of the focal interests of the present thesis is to address this question, utilizing process measures such as eye-tracking technology and stimulated recall protocols.

V. How to Measure Cognitive Processes

The review of previous literature on L2 reading has revealed a common methodological issue, i.e., the need to explore learners' internal processes. More specifically, exploring learners' cognitive processes during reading enables researchers to document how L2 readers regulate and adapt their reading processes in order to achieve different task goals. Research into cognitive processes can also provide valuable information with respect to whether target L2 constructions are attended to or noticed by learners. In a similar vein, in the field of TBLT, Révész (2014) proposes that cognitive processes triggered by task demands need to be documented in order to examine how key conceptualizations associated with task-based instruction, most notably noticing, operate during task performance. Against this background, this section discusses the usefulness of verbal reports (Ericsson & Simon, 1993) and eye-tracking technology (Rayner, 1998), which have been proposed as powerful research tools for exploring learners' cognitive processes.

1. Verbal reports

Verbal reports can be categorized based on the temporal frames in which they are collected. Concurrent reports, e.g., think-alouds and note-taking, are collected as learners verbalize their thinking processes while simultaneously performing the task, whereas retrospective reports, e.g., questionnaires, stimulated recall, interviews and diary entries, are collected when learners verbalize after completing the task (Bowles, 2008, 2010). Verbal reports can also be categorized based on the level of detail. Verbal reports in which learners verbalize their thoughts per se are referred to as *nonmetalinguistic,* and those in which learners provide explanations and justifications

are categorized as *metalinguistic*. In SLA research, verbal reports are generally used to obtain information about learners' internal processes while interacting with the L2 (e.g., Egi, 2004; Mackey, Gass, & McDonough, 2000; Rosa & Leow, 2004; Rosa & O'Neil, 1999). In the field of language testing, verbal reports have played an important role in validating assessment instruments and offered a way to gather more direct evidence that supports researchers' judgments (e.g., Cohen & Upton, 2007; Green, 1998).

Yet, the inherent limitations in use of verbal protocols remain nonetheless. First, as learners cannot report everything that they notice, what learners do not verbally report cannot be documented (Jourdenais, 2001; Rosa & O'Neill, 1999). Second, in the case of retrospective protocols, the veridicality (i.e., veracity and accuracy) of learners' reports and memory decay can pose potential threats (Egi, Adams, & Nuevo, 2013; Ericsson & Simon, 1993; Leow, 2000; Leow & Morgan-Short, 2004; Leow & Hama, 2013). Third, in the case of concurrent protocols, verbal reporting while simultaneously performing the learning task may force learners to engage in dual tasks and thereby interfere with their learning processes (i.e., *negative reactivity*) or, inversely, facilitate learners to perform better (i.e., *positive reactivity*) (Bowles, 2010; Goo, 2010). Fourth, the inherent nature of verbal reports makes them more suitable for qualitative analysis and prone to create room for inconsistent interpretations of collected data. Lastly, it is more likely that the data represent learners' conscious experience rather than underlying cognitive processes, such as attention (Godfroid, Boers, & Housen, 2013; Winke, 2013), due to issues with the "reportability" (Dehaene & Changeux, 2004, p. 1145) of the processes in focus.

As for concurrent reports, Ericsson and Simon (1993) suggest that nonmetalinguistic verbalizations could be expected to be nonreactive, whereas metalinguistic verbalizations might be more reactive. Indeed, previous studies have shown that nonmetalinguistic reports are in general nonreactive, having limited

influence on the comprehension or learning of target form-meaning mappings (Bowles, 2008; Bowles & Leow, 2005; Leow & Morgan-Short, 2004), whereas metalinguistic verbalizations, which require learners to report additional information that might not otherwise have been produced, were more likely to affect learning (Bowles, 2008). For example, Leow, Hsieh, and Moreno (2008) conducted a study investigating whether paying attention to target grammatical constructions affected L2 reading comprehension using think-aloud protocols. Seventy-two university students were randomly assigned to four experimental groups (one targeting a Spanish lexical item and three targeting Spanish grammatical items) and one control group. The participants were instructed to read a Spanish text for comprehension, while drawing circles around the targeted Spanish lexical and grammatical items (10 occurrences) and verbalizing their processing. Comprehension was measured using a 10-item multiple-choice test. Analyses of verbal data and comprehension scores demonstrated that there were no differences in the level of comprehension among the five groups, and that learners were able to process the target features for both form and meaning without making extra effort. While the collected verbal data revealed valuable information regarding learners' simultaneous attention to form and meaning and its impact on comprehension, whether concurrent verbalizing interfered with the participants' natural task performance remains open to debate.

In order to inspect potential reactivity of think-alouds in Leow et al.'s (2008) study, Morgan-Short and her colleagues (Morgan-Short, Heil, Botero-Moriarty, & Ebert 2012) conducted a replication study, adopting the text, the target constructions, and the reading comprehension items directly from the original study. In this research, 205 university students were randomly assigned into either think-aloud or non-think-aloud mode. The results showed that think-alouds had a significant reactive effect on comprehension scores, while the amount of variance explained by group assignment

was only 1%. The researchers interpreted this number as practically not meaningful (Ferguson, 2009), and conclude that think-alouds did not appear to compromise internal validity. It was also found that the deeper the level of processing, the higher the comprehension scores. As suggested by Leow et al. (2008), Morgan-Short et al. (2010) asserted that in the written mode, learners seemed to be better able to attend to both form and meaning unlike in the aural mode (cf. VanPatten, 1990), and that "the cognitive constraints may differ when processing aural versus written L2 input" (p. 679). It should be noted, however, backtracking or rereading behaviors were not controlled in this study, which could have been better identified and addressed with different methodology such as eye-movement recording.

The issue of reactivity may not be a major concern in the case of retrospective verbal reports, as participants do not experience the dual burden imposed by concurrent verbalizations. Yet, the extra learning opportunity that arises from a second exposure to stimuli and verbalizations of cognitive processes can potentially boost the likelihood of detecting learning effects of treatments (i.e., positive reactivity) (Gass & Mackey, 2000). Indeed, Adams's (2003) study revealed that stimulated recall after performing a collaborative writing task while receiving feedback significantly enhanced the accuracy of subsequent writing. By contrast, Egi's (2008) study did not show reactivity of stimulated recall in posttest scores. More specifically, the study included three groups: the first group produced stimulated recall based on video clips, the second group watched the video clips without verbalization, and the third group was a control group. The results of a posttest following stimulated recall sessions revealed no significant impact of either the stimulated recall or the stimulus. In a more recent study, Egi, Adams, and Nuevo (2013) examined whether learning was influenced by non-metalinguistic stimulated recall and metalinguistic stimulated recall. Participants were 29 learners of English, who were randomly assigned to a non-metalinguistic stimulated

recall group, a metalinguistic stimulated group or a control group. The stimuli were video clips recorded during teacher-fronted classroom interactions. The results showed that there were no significant differences among the groups, indicating non-reactivity.

Research into testing the veridicality of verbal reports has also been growing. As for retrospective reports, the threat can be avoided by reducing the time delay between task performance and verbalization. Also, providing learners with some stimulus, such as audio- or videotape of their performance, can minimize the potential threat of veridicality (e.g., Adams, 2003; Egi, 2004, 2008; Egi et al., 2013; Mackey, 2006). For example, Philp and Iwashita (2013) used video-stimulated recall to investigate whether engaging in interaction tasks and observing others interact differentially affected learners' awareness of the target language. In this study, evidence of noticing was operationalised as the participants' "articulation of response to the input, or to their own output, indicative of a perception of form, without distinguishing the degree of understanding involved" (Philp & Iwashita, 2013, p.358), as evinced in their stimulated recall protocols. The results revealed that, in general, more evidence of noticing was found in the Interactors group than in the Observers group. Based on these findings, Philp and Iwashita suggested that engaging in production tasks, even when feedback is not provided, could be beneficial to language learning.

Verbal protocols have long been supported as a promising tool for examining the validity of reading tests (Cohen, 1994; Cohen & Upton, 2007; Goa & Gu, 2008; Green, 1998; Phakiti, 2003; Rupp, Ferne & Choi, 2006; Yamashita, 2003). For example, Green (1998) comments, "[verbal protocols] offer a means for more directly gathering evidence that supports judgments regarding validity than some of the other more quantitative methods" (p. 3). Based on this recognition, Cohen and Upton (2007) conducted a qualitative study in order to test the construct and external validity (the extent to which inferences made in a study can be generalized) of a new format for the

TOEFL reading section. More specifically, it was examined whether learners used different strategies when performing on the traditional single-selection, multiple-choice format versus the new multiple-selection, drag-and-drop format. The latter format was included in TOEFL to simulate academic skills. The participants of this study were 32 learners of English, and they performed the test in either the traditional format or the new format while verbalizing their introspective processes. Overall, the results showed that the respondents employed academic reading-like abilities in order to successfully perform in the new test format. While this study showed learners did indeed employ different test-taking strategies for different test formats, it should be also noted that what the participants did not report could not be analysed. Also, verbalizing concurrently while working on the tests could have led the participants to respond to the test items more conscientiously, as well as in a more structured way, than they would in normal test-taking circumstances. Moreover, as Cohen and Upton acknowledged, a distinction was not made between the strategies used for items that were answered correctly versus incorrectly. A closer inspection of this variable might paint an interesting picture of which strategy learners adopt when coping with test items about which they are uncertain.

2. Eye-movement data

Eye-movement recording, colloquially referred to as eye-tracking, has recently received increasing attention among SLA researchers. Roberts and Siyanova-Chanturia (2013), for example, propose that eye-tracking can serve as a valuable research tool for gathering and exploring the visual mechanics behind learners' real-time processes, as "it allows for the study of moment-by-moment processing decisions during natural, uninterrupted comprehension, and critically, without the need to rely on participants' strategic or metalinguistic responses" (p. 214). Thus, eye-movement data can be useful

when examining in which task condition learners are more likely to attend to grammatical details in the input and how individual differences might come into play.

Rayner's (1998) review of research into reading, although primarily focused on native English speakers' *default* mode of reading, provides useful insights into how eye-movement data can help us understand L2 learners' reading processes. Eye-movement behaviour consists of *eye fixations*, during which the eye dwells on part of a text and processes the incoming input, and *saccades* that occur when the eye moves from one location to the next. The time in-between two saccades is referred to as the *fixation duration*, and this is influenced by both low-level (i.e., visual) and high-level (i.e., cognitive) factors. When investigating fixation durations, early and late processes of reading are often distinguished and examined separately. Early measures are deemed to be sensitive to processes associated with first-time reading, such as word recognition, lexical access and the spontaneous integration of incoming information, while late measures are believed to be sensitive to processes related to the comprehension of text, such as reanalysis, discourse integration, revisits and recovery from processing difficulties (Rayner, 1998; Roberts & Siyanova-Chanturia, 2013; Winke, Godfroid, & Gass, 2013). When readers attempt to correct their inefficient text processing, they tend to exhibit the following types of eye-movements: (a) *regression*, backwards motion for a distance of a few letters, (b) *return sweep*, returning to a precise fixation point, (c) *backtrack*, moving back to discover the source of difficulty, and (d) *corrective saccade*, re-identifying text (Rayner, 1998). In sum, eye-movement data can be valuable, providing a multidimensional and multifaceted picture of text processing during reading.

Eye-tracking technology has been used in the field of L2 sentence processing (e.g., Alptekin & Erçetin, 2015; Dussias, Kroff, Tamargo, & Gerfen, 2013; Jackson, Dussias, & Hristova, 2012; Kaushanskaya & Marian, 2007; Keating, 2009; Siyanova-Chanturia, Conklin, & Schmitt, 2011), mostly in order to explore how learners process L2

grammatical morphemes. For example, Dussias et al. (2013) investigated whether grammatical gender facilitates noun recognition during L2 sentence processing. Sixteen L1 Spanish speakers (control group) and 18 L1 English speakers learning Spanish (high and low proficiency) were provided with two picture images in which genders matched or did not match. While viewing the images, the participants listened to sentences in which articles preceding the target nouns agreed in gender with one or both of the pictures. Another 15 L1 Italian speakers learning Spanish were additionally tested to investigate if the presence of gender in L1 modulated processing of L2 gender marking. All participants' eye-fixations were recorded while performing the experimental tasks. An analysis of their eye-movements revealed that the L1 Spanish speakers and the highly proficient English-Spanish speakers looked at the target items sooner when the two pictures belonged to different gender classes. The less proficient English-Spanish speakers, however, did not use gender information when processing sentences. Italian-Spanish bilinguals also exhibited gender anticipatory effects, but only for feminine items. In short, the results of this study indicate that L2 proficiency and L1-L2 similarity might affect the processing of morphosyntactic information during speech processing.

SLA researchers have also increasingly used eye-movement data to examine whether more attention or noticing leads to more L2 learning (Godfroid, Housen, & Boers, 2010; Godfroid, Boers, & Housen, 2013; Godfroid, Winke, & Gass, 2013; Godfroid & Uggen, 2013; Indrarathne & Kormos, 2016; Morgan-Short et al., 2013; Pellicer-Sánchez, 2016; Roberts, 2012; Sagarra & Ellis, 2013; Smith, 2012; Winke, 2013; Winke, Gass, & Sydorenko, 2013). The underlying assumption is that cognitive engagement with input (e.g., cognitive registration and further processing of input) will mentally take time, resulting in longer eye fixations on linguistic elements (Godfroid, Boers & Housen, 2013). Simply put, learner's eye-movement, i.e., overt attention, is

viewed as manifestation of his or her mental focus, i.e., covert attention, during performing a task. For example, Godfroid and Uggen (2013) used eye-tracking technology to investigate whether learners paid more attention to irregular verb morphology during sentence processing. Forty beginning-level learners of German were provided with twelve German sentence pairs containing stem-changing verbs, and another twelve pairs containing regular verbs. An eye-tracker recorded the participants' eye-movements while they were processing the sentence pairs. A cued-production test was used as a pre- and posttest to measure the participants' knowledge of German stem-changing verbs. The results from eye-movement data and scores from a posttest revealed that the participants looked at stem-changing verbs longer than regular verbs. Also, longer eye-fixation on stem-changing verbs was shown to have a favourable effect on scores in a subsequent production posttest. Based on these findings, Godfroid and Uggen suggested that it is possible that beginning learners attend to irregular morphological features during sentence processing and that the amount of attention paid to stem-changing verbs might relate positively to acquisition of the form-meaning mapping.

This study demonstrates that eye-movement provides highly objective and sensitive information about L2 cognitive processes. As mentioned earlier, verbal reports have been the typical choice of researchers when investigating learners' attention to targeted constructions (Leow, 1997a, 2001b; Rosa & Leow, 2004; Rosa & O'Neil, 1999). However, when relying on verbal reports, "researchers are using a test's sensitivity in measuring a construct to define the construct itself" (Winke, 2013, p. 239). By contrast, eye-tracking technology enables researchers to ensure comprehensive documentation of real-time eye-movement, allowing the obtaining of a fuller and more accurate picture of various cognitive processes such as attention and noticing. Moreover, eye-movement data can capture the amount of attention paid to a certain part of a text as

well as possible comprehension breakdowns and remedial processes to recover from it. With respect to the usefulness of eye-tracking technology, Godfroid et al. (2013) aptly encapsulate that this method (a) reflects accurate and objective information on the dynamicity of cognitive processes in real time by showing both spatial and temporal eye-movement, (b) does not interfere with the primary task (i.e., reading) and hence is nonreactive, and (c) enables quantitative analysis, which is valuable in resolving theoretical issues regarding the role of attention, noticing and awareness in SLA. They also note, however, that eye-movement data are quantitative, and thus may not display the quality of attention. In other words, eye-movement data do not provide information about the depth of cognitive processing, the amount of mental effort or the level of awareness, on which the robustness of the memory trace depends (Godfroid, Housen & Boers, 2010; Godfroid, Boers, & Housen, 2013; Winke, 2013). For this reason, it is highly recommended to triangulate eye-movement data with additional qualitative data to build a complete understanding of the learner's internal reading and learning processes.

In the field of language testing, eye-movement data can also be used to examine the cognitive validity (the extent to which a reading task triggers the type of cognitive processes purported by the task designer) (Bax, 2013; Bax & Weir, 2012; Brunfaut & McCray, 2015) of a reading test. For example, Bax (2013) investigated test-takers' differential reading behaviours and strategy use while completing IELTS reading test items. More specifically, Bax examined whether the test items, i.e., sentence completion and matching, indeed induced test-takers to engage in targeted types of reading, i.e., careful local reading and expeditious local reading. Eye-movement data were collected during reading, and stimulated recall interview data were collected from both successful and unsuccessful participants. The results revealed that successful and unsuccessful learners underwent significantly different processes at various levels, from lexical to

grammatical processing. Also, eye-movement data revealed exactly which parts of the text and which test items were more useful in distinguishing successful readers from unsuccessful readers, even at local-level comprehension. The results led Bax to champion the usefulness of the eye-tracking method for testing the cognitive validity of a reading test when combined with verbal reports.

More recently, Brunfaut and McCray (2015) also used eye-tracking technology to examine the reading processes of 25 test-takers while performing four types of reading tasks in an Aptis test, which included a multiple-choice gap-filling task, a sentence reordering task, a banked gap-filling task and a matching headings task. Each task was targeting different level bands, i.e., A1, A2, B1 and B2, as delineated in the Common European Framework of Reference for Languages (CEFR) (British Council, 2014). Test-takers' eye-movements were recorded while performing the four types of assessment tasks, and stimulated recall protocols were obtained while viewing recorded eye-movements for data triangulation, which generated a rich data set. The results revealed that global processing measures, such as total fixation time and total number of fixations, and local processing measures, such as the number of forward saccades and the number of regressions, increased as the CEFR level of the task increased. Stimulated recall protocols further revealed that A1 (multiple-choice gap-filling task) and B1 (banked gap-filling task) depended on lower-level processes, such as syntactic parsing or lexical access, extensively, whereas A2 (sentence reordering task) and B2 (matching headings task) were more closely linked to higher-level processes, such as building a global representation of the text. It should be noted, however, that each CEFR level included a distinct task type, and hence the different reading processes found in this study might also reflect the influence of task type rather than CEFR levels.

With respect to more practical concerns, Spinner, Gass, and Behney (2013) suggested that clear standards for methodology should be established, as their empirical

study showed that small changes in font size, textual arrangement and other display features brought about significant differences in the way learners respond to linguistic elements in a given text. The researchers compared two display conditions in which learners of Italian were asked to process Italian gender-marking. In experiment 1, the article and noun were presented on separate lines in a larger font, whereas in experiment 2, the article and noun were presented on the same line in a smaller font, closely aligned to normal reading material (i.e., an ecologically more valid condition). The eye-movement data revealed that, in experiment 1, an equal amount of time was spent on each noun and article, implying that learners employed both morphophonological and morphosyntactic cues when processing article-noun agreement; on the other hand, in experiment 2, nouns and articles were captured within one gaze, resulting in less time being spent on articles. Based on these findings, Spinner et al. suggested that (a) small changes in display features might bring about large changes in eye-movement data and (b) breaking texts up over several lines in larger fonts is recommended to better capture learners' form-meaning processing, rather than mimicking natural reading conditions.

3. Summary

This section has covered research methodologies available to detect and document learners' internal processes during reading. In particular, verbal reports and eye-movement data were explored as means to collect information about learners' cognitive processes during reading. The data obtained using these methods will shed light on the modulating effects of tasks on L2 reading processes as well as the noticing of target constructions during reading. In addition, several limitations associated with these methods were identified and discussed, such as veridicality and reactivity in relation to verbal reports and the importance of clear task layout for the accurate capture of learners' eye-movements.

VI. Working Memory Capacity

Over the past two decades, working memory has attracted attention from L2 researchers regarding how individual differences in working memory capacity account for differential performances in cognitively complex tasks, such as reading comprehension and language-learning. This section presents an overview of the cognitive architecture of working memory and empirical studies that have investigated its relation to L2 reading and L2 learning.

1. Cognitive architecture of working memory

Working memory capacity has been explored extensively as an important cognitive construct in the field of cognitive psychology, and various theoretical models for working memory capacity have been put forward (Baddeley, 2003a, 2003b; Cowan, 2005; Jarrold & Towse, 2006; Miyake, 2001). The present thesis adopted Baddeley's (2007) framework, considering its substantial impact on studies on the role of working memory capacity in L2 learning. Baddeley and Hitch (1974) provided an outline of the cognitive architecture of working memory, which has been widely used in the fields of cognitive/ educational psychology. According to this framework, working memory consists of *executive control*, a limited attentional control system, and two domain-specific sub-systems, the *phonological loop*, responsible for temporary storage of verbal and acoustic information, and the *visuo-spatial sketchpad*, specialized in storing and processing visuo-spatial information (see Figure 5).

The two working memory components that are particularly relevant to the present thesis are phonological loop and the control executive. The phonological loop subsumes two subcomponents, i.e., a temporary storage system that holds phonological information for a few seconds until it decays, while a subvocal rehearsal system maintains and registers the information stored in short-term memory. As the phonological loop in particular pertains to the retention of sequential information, its

function is typically measured with tasks that require the immediate repetition of sequences of digits or words/ nonwords in the order of presentation (Baddeley, 2000). When a sequence of letters is presented, their retention and recall depend on their phonological qualities. For instance, Baddeley (1966a, 1966b) showed that sequences of similar sounding letters or words, such as *man, cat, map, cab, can,* were correctly recalled at a success rate of less than 20 per cent, whereas dissimilar sequences, such as *pit, day, cow, sup, pen,* were correctly recalled with a success rate above 80 per cent. Evidence for the subvocal rehearsal system also comes from serial recall studies using words of varied length. In Baddeley, Thomson, and Buchanan's (1975) study, for example, multisyllabic words were recalled significantly more poorly compared to monosyllabic words.

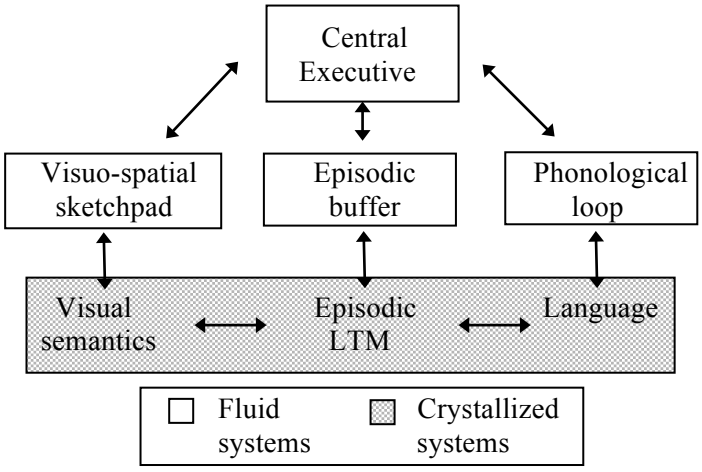

Figure 5. Current multi-component model of working memory (*Source*: Adopted from Baddeley, 2003a, p.203)

Executive control is responsible for attentional control, carrying out conscious processing, monitoring, intentional learning and searching for solutions to problems. Thus, executive control is deemed to be the principal factor responsible for individual differences in complex working memory span. Complex working memory span is frequently measured by requiring learners to read out a series of sentences while remembering the last word in each sentence for immediate recall. Performance on such tasks has proven to be a robust predictor of diverse complex cognitive abilities such as

reading comprehension (Daneman & Carpenter, 1980). Baddeley (2000) recently added a memory sub-system under executive control, the *episodic buffer*, where integrated processing outcomes are stored and smaller chunks of information become larger units, while activating relevant information from long-term memory. Thus, while the episodic buffer is limited in its capacity, drawing on executive control, it also differs from executive control in that its major function is not attentional control but creating a single multi-faceted code through combining information from different modalities.

When it comes to L2 learning, working memory capacity has been discussed as a plausible and stable component of language aptitude, explaining varying levels of ultimate attainment in adults' L2 acquisition (Kormos, 2013; Miyake & Friedman, 1998; Robinson, 1995a, 2005a; Sawyer & Ranta, 2001; Skehan, 2002, 2009; Yalçin, Çeçen, & Erçetin, 2016). As Sawyer and Ranta (2001) suggest, provided that a certain level of attention or noticing is crucial to L2 learning, and that attention at any moment is limited by working memory capacity, it appears logical to posit a relationship between learning and working memory capacity. Indeed, given that working memory serves as an arena where pieces of information are identified, selected, analysed and integrated, there will be few, if any, attentional resources left for learning if working memory is overloaded. Given that the focus of this paper lies in L2 learning through L2 reading, a review of previous studies that have explored the role of working memory in L2 reading comprehension and L2 learning is presented in the following sections.

2. Working memory and L2 reading comprehension

As reviewed above, working memory serves dual functions, i.e., storing information for later retrieval with a brief interval and processing incoming data by means of operational resources (Baddeley & Hitch, 1974). Based on findings from reading span tests, Daneman and Carpenter (1980) proposed that there is a trade-off between storage functions and computation processes, and that a trade-off can be

manifested more markedly in cognitively and linguistically demanding tasks, such as reading comprehension. Indeed, working memory capacity has consistently been shown to correlate strongly with measures of L1 reading comprehension and specific reading skills, which has led researchers to regard working memory as central in explaining individual differences in L1 reading comprehension (Daneman & Carpenter, 1980; Daneman & Green, 1986; Just & Carpenter, 1992; Turner & Engle, 1989).

Harrington and Sawyer (1992) conducted seminal research into the role of working memory capacity in the context of L2 reading. Twenty-four Japanese students learning English as a foreign language completed simple span tests (i.e., a digit span test and a word span test) and a reading span test both in Japanese and L2 English. The index of L2 reading comprehension consisted of scores from a grammar and reading section of paper-based TOEFL and an additional cloze passage. The results of correlational analyses revealed that the participants who scored higher on the L2 reading span test did better on the L2 reading comprehension measures, whereas scores on the L2 digit and word span tests only weakly correlated with L2 reading comprehension. This finding led Harrington and Sawyer to conclude that complex working memory (henceforth, CWM) plays a crucial role in L2 reading comprehension. It seems noteworthy, however, that the L2 reading span test and the L2 reading comprehension test could have measured overlapping constructs, calling for more attention to the domain specificity of the memory measure.

Whether working memory is language-dependent or independent has been studied in Osaka and Osaka (1992) and Osaka, Osaka, and Groner (1993). In Osaka and Osaka's research, L1 Japanese and L2 English participants performed Daneman and Carpenter's (1980) reading span test in L1 and L2 versions. While the procedure was largely the same to that of Harrington and Sawyer (1992), the participants orally recalled the sentence-final words of each set, without a sentence acceptability judgment

task incorporated. The CWM index was the highest number of words that the participants consistently recalled for at least three sets. The results showed that there were significant correlations between L1 and L2 reading span scores, and between Daneman and Carpenter's original version and the Japanese version. In their follow-up research, Osaka et al. (1993) again found a significant correlation between L1 German and L2 French versions of the reading span test. Based on their findings, Osaka et al. suggested that CWM might be independent of any specific language proficiency.

It should be noted, however, that above studies employed Daneman and Carpenter's classic version of the reading span task, which did not measure the processing function of CWM. Regarding this issue, Waters and Caplan (1996) argue that it is important that semantic and syntactic acceptability judgments be included as part of a reading span test and reaction time should also be recorded so that any trade-off between storage and processing can be taken into account. In Waters and Caplan's version of the reading span test, CWM scores were obtained using three measures: (a) mean reaction time taken for correct responses on the acceptability judgments, (b) number of errors in an acceptability judgment, and (c) number of trials where sentence-final words were incorrectly recalled. Correlation analyses showed that reaction time correlated negatively with recall errors and judgment errors, indicating a trade-off between storage and processing. Based on these findings, Waters and Caplan argue that participants "with the same working memory span as measured by the recall component of a complex span task may be performing very differently on the processing component of the task" (p. 64). This procedure was adopted in more recent research investigating the role of CWM in L2 reading comprehension (e.g., Alptekin & Erçetin, 2009; Leeser, 2007; Walter, 2004).

While the above studies have shown a possible interplay between CWM and L2 proficiency in L2 reading, there are also studies that focused on topic familiarity as a

moderator of the effects of CWM on L2 reading comprehension (e.g., Alptekin & Erçetin, 2009, 2011; Leeser, 2007). For example, Leeser (2007) investigated the joint influence of working memory capacity and topic familiarity on L2 reading comprehension and the processing of Spanish future tense morphology embedded in texts. In this study, 146 beginning L2 Spanish learners completed a questionnaire rating their familiarity with ten topics, four of which were the topics of the reading text. L2 reading comprehension was measured using a written recall task completed in the L1 English. CWM was measured via a computerized version of Waters and Caplan's (1996) reading span test. Knowledge of the target feature was measured with a form recognition task and a form production task. The results showed that CWM scores had a significant effect on performance on the written recall task and the form recognition task, but only in the familiar condition. Leeser suggested that individuals with a higher CWM might be able to conjure up more domain knowledge during reading than those with lower CWM.

Similar results were found in Alptekin and Erçetin's (2009, 2011) studies. They explored whether L2 CWM played differential roles for literal and inferential comprehension (Kintsch, 1998). Thirty Turkish learners of Engliss carried out two versions of L2 reading span tests: one version included a recall task and the other version involved a recognition task. A narrative text was selected from an American short story, and reading comprehension was measured using a multiple-choice test, half of which was to measure literal understanding and the other half inferential understanding of the text. The results from correlational analyses showed that CWM scores emerged as a significant predictor of L2 reading comprehension scores, but only for the inferential level and when the storage component was measured through a recall-based procedure. With respect to this finding, Alptekin and Erçetin claimed that "inferential bridging and elaboration, on their own, place heavier demands on WM as a

result of the intrinsic complexity of the tasks they involve" (p. 258). They also suggested that there would be larger variance in CWM scores when measured through a recall task, as participants have to evoke internally generated cues to delimit the search set, which will tax their limited CWM.

More recently, Alptekin, Erçetin, and Özemir (2014) further attempted to investigate whether the relationship between working memory and L2 reading comprehension ability was moderated by the language used in a reading span test (L1 vs L2) and the type of processing task in the reading span test (semantic vs morphosyntactic anomalies). Ninety-eight Turkish undergraduate students completed four versions of reading span tests adapted from Daneman and Carpenter's (1980) test in both L1 and L2. The reading span tests asked participants to judge if there were anomalies, semantic or morphosyntactic. Participants'' L2 reading comprehension ability was measured with multiple-choice questions. The results from exploratory factor analysis revealed that the storage scores of reading span tests shared a common underlying factor, implying that a storage component might be language- and task-independent. By contrast, the processing scores were affected by the language and type of task. As for L2 reading comprehension scores, significant correlations were found with storage scores when determining semantic anomalies in either L1 or L2, and with processing scores that entailed judging semantic anomalies in L1 or morphosyntactic anomalies in L2. Based on these findings, the researchers suggested that the language and the type of task used in a reading span test should be considered carefully, as they may confound the relationship between working memory and L2 reading comprehension.

In sum, previous studies have revealed that working memory plays an important role in L2 reading comprehension. As Waters and Caplan (1996) point out, to better evaluate working memory capacity, process measures, such as reaction time on

acceptability judgments, should be taken into account. Also, as Alptekin and Erçetin (2009) suggested, the storage component of working memory may be better assessed via recall tasks than through recognition tasks, magnifying the variance in learners' individual differences and hence increasing the effect size. Next, while it is generally acknowledged that working memory capacity is a domain-general ability (Osaka & Osaka, 1992; Osaka et al., 1993), there is also some evidence that working memory may interact with L2 proficiency (Walter, 2004), which warrants further research. In addition, given that the propositional text model and the situation model are built through interconnected but distinct cognitive processes (Kintsch, 1998), more research seems to be needed on how working memory plays different roles in establishing literal and inferential comprehension (Alptekin & Erçetin, 2009, 2011). Finally, Leeser's (2007) study revealed a potential interplay between working memory capacity and topic familiarity, which highlights the need to control learners' prior knowledge about a topic when researching the role of working memory in L2 reading.

3. Working memory and L2 learning

The role of working memory in L2 learning has been of particular interest among researchers who take a usage-based approach to language learning (e.g., N. Ellis, 1996, 2005; N. Ellis & Sinclair 1996; Williams & Lovatt, 2005). In this view, language learning is equated with sequence learning, which is a memory-driven process wherein rules emerge from implicit analysis and generalization of morpheme sequences stored in long-term memory. Phonological short-term memory (henceforth, PSTM) is deemed to be responsible for storing and processing morphemes, and PSTM is seen as playing an instrumental role in the acquisition of vocabulary and morphosyntactic development (N. Ellis, 1996, 2005; N. Ellis & Schmidt, 1997).

In order to test this theoretical assumption, N. Ellis and Sinclair (1996) investigated if the articulatory rehearsal of L2 utterances would enhance 87 university

students' learning of Welsh words and grammatical structures. The participants had no experience of Welsh in advance of the study, and they were randomly assigned to one of three groups: (a) a repetition group, in which they were instructed to repeat Welsh utterances every time, (b) an articulatory suppression group, in which they were prevented from articulating Welsh utterances, and (c) a control group, with no instruction given. The learning materials consisted of 30 Welsh utterances, ten of which contained soft mutations of Welsh. In each trial of the learning phase, the computer played pre-recorded utterances, and the participants were instructed to respond to them by typing in the corresponding English translations. After the learning phase, a timed grammaticality judgment test, a metalinguistic awareness test and a speech production test were administered to measure the participants' knowledge of Welsh. The scores from the three tests revealed that phonological rehearsal of Welsh utterances resulted in significantly more gains in learning of the target Welsh words and phrases, in addition to grammatical regularities including Welsh soft mutation. N. Ellis and Sinclair concluded that repetition of L2 utterances stimulated short-term phonological storage as well as eventual establishment of long-term sequence formation, resulting in greater gains in acquiring the lexical and grammatical rules underlying the utterances. Yet, it should be noted that this study did not include an independent measure of the participants' PSTM, which renders their conclusion only speculative. In addition, as N. Ellis and Sinclair admitted, the articulating of Welsh utterances (a) could have resulted in hearing their own utterances, which doubled the amount of input, and at the same time (b) served as an output production activity, which enhanced the level of attention to the Welsh input, which might have confounded the results of the study.

Williams and Lovatt (2003) developed separate measures of PSTM in order to investigate the relationship between phonological memory and learning a semi-artificial language. The participants were provided with 32 semi-artificial lexical items

containing rules for determiner-noun agreement, plural constructions, and masculine and feminine inflections, and completed a learning phase that consisted of a fragment completion task and a translation task. PSTM was tested via a serial recall task in which the target nouns, in either singular or plural forms, were used as the stimuli. Additionally, the participants performed a morpheme-combination memory task and an input memory task. Learning was measured through a rule-learning test comprising a production task and an English translation task with novel words. The results revealed a significant correlation between PSTM scores and the ability to acquire the abstract categorization of nouns into word classes. Based on these results, William and Lovatt supported the role of PSTM, even in the learning of fairly abstract aspects of L2 grammar. Yet, caution is warranted as the learning tasks used in this study, i.e., a fragment completion task and a translation task, resembled rote memorization accompanied by pushed output, and hence seemed to be fairly explicit. That said, the findings of this study might not be readily generalizable to natural learning situations wherein L2 is learned largely through implicit processes.

As PSTM is often measured with serial recall tasks that require learners to repeat series of words or nonwords, the possible influence of vocabulary knowledge on PSTM has continually been pinpointed by researchers (e.g., Engel de Abreu & Gathercole, 2012; French, 2006; Service, 1992; Service & Kohonen, 1995). In other words, previous studies on the association between PSTM and L2 grammar learning have repeatedly found that the relationship is mediated by participants' L2 vocabulary knowledge, which implies that PSTM is not independent of language learning but is also influenced by long-term language knowledge. Different results were obtained, however, in French and O'Brian's (2008) study and Verhagen, Leseman, and Messer's (2015) study. That is, even when participants' vocabulary knowledge was controlled, significant correlations were found between PSTM and L2 grammar learning ability:

"individuals with good phonological memory skills are assumed to create more robust and stable phonological representations, which are needed for language learning, than individuals with poorer phonological memory skills" (Verhagen et al., 2015, p. 421).

Kempe and Brooks (2008) focused on the role of CWM in learning rules with differential transparency. In experiment 1, 43 adults were presented with a subpart of the Russian case-marking system in which the gender of nouns was transparently marked in the nominative case (12 masculine, 12 feminine). In experiment 2, 44 participants were presented with a similar system of nouns, but with a non-transparent gender-marking rule in the nominative case. The CWM of the participants was measured with Daneman and Carpenter's (1980) reading span task, and nonverbal intelligence was assessed using the Cattell Culture Fair Test (Cattell, 1971). Both learning sessions and assessment tasks included describing pictures of various objects after listening to dialogues about the pictures. Also, a vocabulary test was administered wherein the participants were presented with pictures of objects, one at a time, and asked to name the object in each picture. The results revealed that learning was more successful in experiment 1 than experiment 2. Also, along with nonverbal intelligence, CWM scores were shown to have a positive impact on learning the transparent system, but not the nontransparent system. The results led Kempe and Brooks to conclude that adult learners benefited from regularity when learning morphological patterns, but not necessarily the underlying rules, due to the complexity of morphological variations, which led learners to depend, inevitably, on item-based learning.

Martin and N. Ellis (2012) addressed both PSTM and CWM and their associations with learning the vocabulary and grammar of an artificial language. Forty university students completed a nonword repetition task (Gathercole, Pickering, Hall, & Peaker, 2001) and a nonword recognition task (O'Brien, Segalowitz, Collentine, & Freed, 2006), while their CWM was measured with a listening span test adopted from Harrington and

Sawyer (1992). The artificial language was taught in three one-hour sessions, accompanied by English translations and relevant pictures. After the learning sessions, a translation task and a recognition task were administered to measure the amount of learning. The results revealed a significant correlation between PSTM scores and vocabulary test performance, as well as between CWM and grammar scores. Based on these results, Martin and N. Ellis suggested that, compared to learning vocabulary, learning grammatical patterns requires more global processing, including storing a greater amount of information over time and analysing relevant information stored in long-term memory. It should be noted, however, that as in Williams and Lovatt's (2003) study, this study entailed an artificial laboratory language, which imposes inherent limitations on accounting for real-life language learning.

The differential contributions made by PSTM and CWM to L2 learning were also found in Kormos and Sáfár's (2008) study that included learners with different L2 proficiency levels. The participants were 121 secondary school students aged 15 to 16. Twenty of them were of pre-intermediate level while the others were at a beginners' level. Their PSTM and CWM were measured with a nonword repetition task and a backward digit span task, respectively (Racsmány et al., 2005). Their English proficiency was measured with a Cambridge First Certificate Exam, which consisted of speaking, listening, writing, reading and use of English (vocabulary and grammar), at the end of two consecutive years. The results from correlational analyses revealed that, for the beginning learners, CWM scores correlated strongly with overall language proficiency test scores. In the case of pre-intermediate learners, in contrast, PSTM scores shared a significant correlation with the whole language proficiency test. Kormos and Sáfár suggested that the different learning situations could explain the discrepancy. In this study, beginning learners were situated in a relatively explicit learning context, which required the memorization of rules and their applications, which might explain

why CWM played an important role in their success. In contrast, in the case of pre-intermediate learners, learning was mostly implicit, which might explain the significant correlation between PSTM and language proficiency scores (e.g., Masoura & Gathercole, 2005).

The role of working memory capacity in L2 learning has also been researched within the interactionist framework, focusing on the efficacy of corrective feedback, especially that of recasts (e.g., Baralt, 2010; Goo, 2012; Li, 2013; Mackey, Philp, Egi, Fujii, & Tatsumi, 2002; Révész, 2012; Sagarra, 2007; Sagarra & Abbuhl, 2013; Trofimovich et al., 2007; Yilmaz, 2013). Some studies have included both PSTM and CWM measures (e.g., Mackey et al., 2002; Révész, 2012; Segarra, 2007), while most studies have focused on CWM. Also, recasts have received notable attention, while Goo (2012), Li (2013) and Yilmaz (2013) compared the moderating role of working memory on the efficacy of recasts and other types of corrective feedback. Some studies have revealed significant correlations between working memory capacity and the effects of recasts on the development of L2 target feature(s) (e.g., Goo, 2012; Mackey et al., 2002; Révész, 2012; Segarra, 2007; Yilmaz, 2013), whereas other studies have found only limited relations (e.g., Baralt, 2010; Trofimovich et al., 2007). Also, the results from Révész's study illustrated that CWM scores correlated with written posttest scores whereas PSTM scores correlated with oral posttest scores. Based on these findings, she suggested that PSTM might play a bigger role in the development of procedural knowledge by enabling learners to maintain the information in recasts for longer, resulting in more robust long-term memory traces. CWM, on the other hand, might play an important role in the development of declarative knowledge or performing literacy skills whereby required learners to hold verbal information in PSTM while processing other cognitive activities. Lastly, it seems noteworthy that, in Baralt's study, as

reviewed in an earlier section of this paper, working memory was shown to mediate the efficacy of recasts only in face-to-face mode, but not in computer-mediated mode.

In sum, PSTM and CWM have been shown to be related to L2 learning, while their relative impacts vary depending on the nature of memory measurement tests (e.g., Martin & N. Ellis, 2012), the tasks used in learning sessions (e.g., Williams & Lovatt, 2003), regularity of the target rule (e.g., Kempe & Brooks, 2008), learners' L2 proficiency and learning situations (e.g., Kormos & Sáfár, 2008), the type of knowledge and literacy skills entailed (e.g., Révész, 2012) and the mode of interaction (e.g., Baralt, 2010). The methodological divergence appears to be a challenge when attempting to elucidate the differential impact of working memory on various aspects of SLA. This review has indicated that language-independent measures (such as L1-based or digit span tasks) seem to reveal a more accurate role for working memory in L2 development by reducing potential covarying effects induced by measuring overlapping constructs (Kempe & Brooks, 2008; Williams & Lovatt, 2003). Also, it seems desirable to employ measures for both PSTM and CWM in order to tap into the multidimensional nature of working memory (e.g., Kempe & Brooks, 2008; Martin & Ellis, 2012; Révész, 2012).

4. Summary

To summarize, working memory, while subsuming multi-components, is limited in its capacity and hence constrains complex cognitive processes. The review of previous studies has, overall, supported the important role of working memory in L2 reading comprehension and L2 learning. The operationalisation and measurement of working memory varied greatly among studies, such as the language used in a memory test (L1 vs L2), measurement method (recall vs recognition; semantic vs syntactic anomalies) and the components involved in a test (storage vs processing), to name but a few. In addition, a wide variety of associated variables emerged, such as learners' L2 proficiency, the level of comprehension entailed (literal vs inferential), and the nature of

the targeted L2 construction (e.g., lexis vs grammar, degrees of regularity and abstractness of the rule). This review provides some useful insights to the present thesis: (a) domain-general/ language-independent tasks should be used to reduce the confounding influence of the language used in memory tests, (b) both PSTM and CWM need to be measured in order to better account for the distinctive contributions made by each to L2 reading and learning, (c) repetition (or recall) tasks, rather than recognition tasks, are more helpful in magnifying the potential variance in learners' performance on working memory measures, (d) learners' topic familiarity to the reading text needs to be controlled, as it may serve as a confounding variable; and (e) as Wen (2012) suggested, a clear demarcation should be made between the two concepts in this thesis: 'working memory' is an umbrella term in its entirety, subsuming all its sub-components, whereas 'complex working memory' refers to the executive control function and phonological short-term memory. These aspects were taken into careful consideration when selecting the working memory measures implemented in the present thesis.

CHAPTER 3

STUDY 1

The review of TBLT literature revealed that the scope of previous research has primarily focused on productive skills, and thus it needs to be expanded into other language skills, such as L2 reading, for a more nuanced understanding of task-based language learning. In addition, research into how task complexity affects L2 learning has generated inconsistent findings, warranting the need for more empirical investigations. To fill these gaps, the following research questions were addressed in Study 1:

RQ (1) To what extent do the cognitive demands of second language reading tasks affect reading comprehension?

RQ (2) To what extent do the cognitive demands of second language reading tasks affect development in the knowledge of target language constructions?

In addition, although much research has looked into the efficacy of glossing in L2 lexical learning, only a few studies have examined its effects on the learning of L2 constructions other than lexis, and so far the findings have been inconclusive (e.g., Guidi, 2009; Martinez-Fernández, 2010; Nagata, 1999). Thus, Study 1 also sought to answer the following research questions:

RQ (3) To what extent does glossing affect second language reading comprehension?

RQ (4) To what extent does glossing affect development in the knowledge of target language constructions?

Lastly, working memory capacity, which has been championed as a central component of both L2 reading comprehension and L2 learning, was included as a potential moderating factor, as expressed in the following research questions:

RQ (5) To what extent does working memory capacity moderate the effects of the cognitive demands of second language reading tasks on reading comprehension?

RQ (6) To what extent does working memory capacity moderate the effects of the cognitive demands of second language reading tasks on development in the knowledge of target language constructions?

RQ (7) To what extent does working memory capacity moderate the effects of glossing on second language reading comprehension?

RQ (8) To what extent does working memory capacity moderate the effects of glossing on development in the knowledge of target language constructions?

This chapter begins with detailed description of the research design, the participants, and the research instruments including treatment tasks, assessment tasks, and questionnaires. Then, the experimental procedure and the rationale for using mixed-effects modelling is presented, followed by the results of data analyses. The chapter concludes with a summary of the results, discussions on the findings, and insights for Study 2.

I. Research Design and Methodology

1. Design

This study examined the impact of two independent variables, task complexity and glossing, on Korean undergraduate students' L2 English reading and learning. As illustrated in Figure 6, the study employed a pretest, posttest and delayed posttest design, with two treatment sessions. Following a 2x2 experimental design, fifty-two participants were randomly assigned to one of four conditions: [+ complex task, + glossing], [– complex task, + glossing], [+ complex task, – glossing] and [– complex

task, – glossing]. Under one of the four task conditions, participants completed two treatment sessions. In each session, participants read a passage taken from a TOEFL exam, while simultaneously answering reading comprehension items. Development in the knowledge of target constructions was measured with a grammaticality judgment test and vocabulary form and meaning recognition tests. Participants' working memory capacity was measured with various span tasks (a digit span task, a backward digit span task, a nonword span task, and an operation span task). Throughout the experiment, different questionnaires were administered to collect information about the participants' English language learning experiences and elicit their reflective responses to the reading tasks. More detailed explanations of the research instruments and procedures are provided in the following sections.

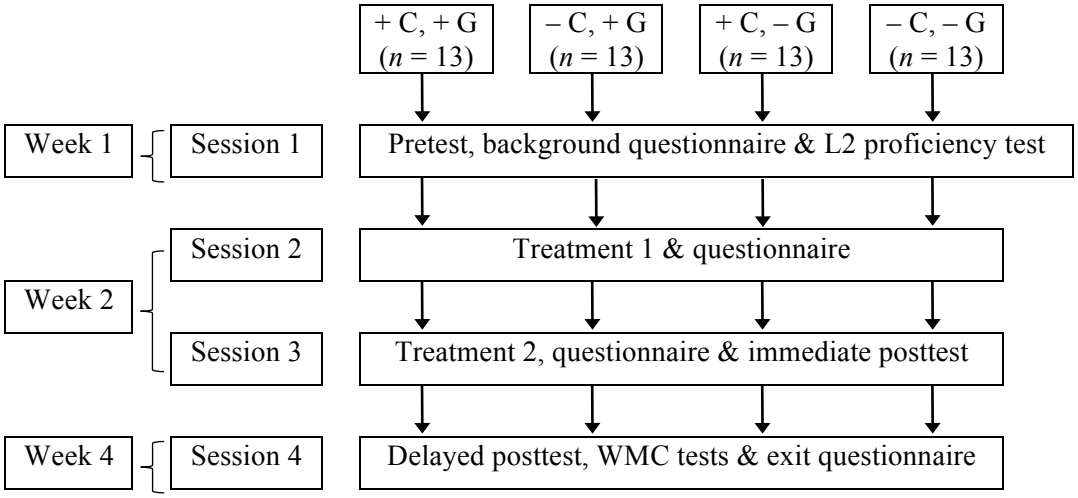

Figure 6. Experimental design and procedure for Study 1

2. Participants

The participants comprised 14 male and 38 female undergraduate students enrolled in a university in Korea. Their L1 was Korean and their average age was 22.84 years (*SD* = 1.94). Their average onset age of English learning was 8.73 years (*SD* = 2.18), and 11 students reported that they had stayed in English speaking countries, such as the US, Australia, Canada, the Philippines and Malaysia (Mean = 6.73 months, *SD* =

4.88 months). They had received no explicit instruction on the target construction (i.e., English unaccusative verbs) prior to this study. To ensure the homogeneity of participants' English ability, their English proficiency was measured with the *Reading and Use of English* section of a practice *Cambridge Proficiency: English* (CPE) test, developed and provided by *University of Cambridge ESOL Examinations*. The target test-takers of CPE are proficient English users, whose CEFR levels range from C1+ to C2. It should be noted that this test turned out to be too challenging for the participants of this study, as evinced from their poor performance on the test in general (for results, see Table 9). Their TOEFL scores reported in the background questionnaire (Mean TOEFL scores = 95.47, *SD* = 10.27) were also shown equivalent to C1 of the CEFR levels. Based on their scores, stratified random sampling was applied in order to reduce sampling error and ensure equivalence among the groups in terms of English proficiency.

3. Materials

3.1. Texts

For this study, two expository texts were selected from passages used for previous TOEFL tests developed by the Educational Testing Service (2013). Both texts were excerpts from university-level textbooks and similar academic materials. The texts were chosen based on two criteria: (a) sufficient numbers of occurrences of the target constructions and (b) unfamiliar topic to the participants. As summarized in Table 4, the titles of the texts were *Petroleum Resources* and *The Cambrian Explosion*. Text 1 explained the formation, extraction and refinement of petroleum resources and the challenges and dangers posed in their use; text 2 reviewed fossil evidence for an evolutionary explosion that happened during the Cambrian period. Text 1 contained 682 words and six paragraphs, whereas Text 2 consisted of 699 words and seven paragraphs. The readability of the two texts was calculated with various indices including *Flesch-*

Kincaid grade level, *Gunning-Fog score*, *Coleman-Liau index*, *SMOG index* and *Automated Readability index,* and the averages were 11.6 for Text 1 and 13.4 for Text 2. Each readability index corresponded to the number of years of education (based on the US education system) required to understand the text. According to their average readability, the texts required an upper-intermediate level of English proficiency and thus were considered appropriate for the participants of this study. Participants' familiarity with the topics of the texts was assessed through post-reading questionnaire items (for results, see Table 12). Additionally, in order to prevent ordering effects, the presentation order of the texts was counter-balanced within each task condition.

Table 4. Characteristics of the treatment texts

	Text 1	**Text 2**
Title	Petroleum resources	The Cambrian explosion
Number of words	682	699
Average readability	11.6	13.4

3.2. Targeted L2 constructions

One target L2 feature of the present study was the English unaccusative construction, which Korean learners have been reported to have persistent difficulty in acquiring (e.g., Hwang, 1999, 2001; J. Kim & H. Kim, 2012; No & Chung, 2006; Oshita, 2000; Shin, 2011). Ten pseudo-words were additionally included in order to examine the effects of task complexity and glossing on the incidental learning of lexical items.

3.2.1. English unaccusative verbs

Perlmutter (1978) first introduced the *Unaccusativity Hypothesis* in which intransitive verbs are classified into either unergatives or unaccusatives.[1] As can be seen in the examples below, whereas an unergative verb assigns an agent role of a volitional

[1] For unaccusatives that have transitive counterparts, some linguists use the term *ergatives* (e.g., Burzio, 1981) or *anticausatives* (e.g., Verrips, 1998). This paper adopts the term unaccusatives, following Perlmutter's (1978) categorization.

act to its subject, the subject of an unaccusative verb lacks volitional control, performing a patient role. Thus, subjects of unaccusative verbs typically undergo a change in state, as in (b) below.

a. Unergative: Mary danced.
   agent
b. Unaccusative: The snow melted.
   patient

SLA researchers have consistently found that L2 learners of English tend to overuse passive structures with unaccusative verbs (Balcom, 1997; Chung, 2014; Croft, 1995; Hwang, 1999, 2001; Ju, 2000; Lee, Miyata, & Ortega, 2008; No & Chung, 2006; Shin, 2011; Oshita, 2000; Zobl, 1989). Overpassivization errors with unaccusative verbs were found to be prevalent, even among high proficiency L2 learners (Zobl, 1989). Following are some examples of overpassivized sentences produced by L2 learners of English.

c. *The most memorable experience of my life was happened 15 years ago. (Arabic: Zobl, 1989)
d. *My mother was died when I was just a baby. (Thai; Zobl, 1989)
e. *First, the change of life-style will be happened. (Korean; Ju, 2000)
f. *You are arrived in the eternity city. (Italian: Oshita, 2000)
g. *Two or three days ago, the important trouble was happened. (Japanese; Oshita, 2000)

Researchers have identified several factors that may affect L2 learners' difficulty in acquiring English unaccusativity. Some researchers (Balcom, 1997; Hwang, 1999, 2001), for example, suggest that if an unaccusative verb has a transitive counterpart, L2 English learners might have a stronger tendency to make overpassivization errors than with non-alternating unaccusative verbs. Below are some examples of alternating and non-alternating unaccusative verbs in English.

h. Alternating unaccusatives: ship, change, close, break
   Non-alternating unaccusatives: happen, result, arrive, disappear (Perlmutter, 1978)

Hwang's (1999, 2001) findings from Korean learners' grammaticality judgment data support this assumption, showing that learners had much more difficulty in acquiring alternating unaccusative verbs than non-alternating ones.

Another source of difficulty in acquiring unaccusative verbs comes from the presence or absence of a conceptualizable agent. Most notably, Ju (2000) provides two unaccusative structures that many L2 learners judge ungrammatical at strikingly different rates: *The car disappeared* (80%) and *The accident happened* (20%). She suggests that the different error rates might be explained by the degree to which an unaccusative verb can have a pragmatically conceptualizable agent. For instance, in the case of *The accident happened*, there is no clear conceptualizable agent, and learners are more likely to accept the sentence as grammatical. In contrast, *The car disappeared* has a pragmatically conceptualizable agent, as cars do not move by themselves. In other words, it lacks the agent responsible for the event, which might lead L2 learners of English to judge this sentence ungrammatical. It should also be noted, however, that Lee et al. (2008) found a non-significant effect of internal/ external causation on Korean learners' grammaticality judgment of English unaccusativity.

While overpassivization of English unaccusative verbs is considered a universal phenomenon, regardless of learners' L1, Oshita's (2000) corpus analysis shows that Korean learners make overpassivization errors to a significantly greater extent (80%) than do Italian and Spanish learners (36% and 26%, respectively). No and Chung (2006) suggest that how the passive voice and unaccusative verbs are expressed in Korean may account for this phenomenon. More specifically, in Korean, passive forms and unaccusative verbs are often realized by same morphological means. Consider the following sentences:

i. *mwun-i* *Kanghee-e* *uyhay* *yel-**i**-ess-ta*. (Passive voice)
door-NOM by open-PASS(*i*)-PST-DC
'The door was opened by Kanghee.'

j. *myun-i* *cecello* *yel-**i**-ess-ta*. (Unaccusative)
door-NOM by itself open-UAC(*i*)-PST-DC
'The door opened by itself.'

(Note: NOM = nominative case, PASS = passive,
UAC = unaccusative, PST = past tense, DC = declarative)

In the above example, the same morpheme *i* is used to mark the passive voice, as in (i), and unaccusativity, as in (j), resulting in an identical form, *yel-**i**-ess-ta*. No and Chung's study found that Korean L2 learners of English were more likely to accept passivized unaccusative verbs as grammatical when the corresponding Korean verbs included passive morphemes (such as *i*, *ci*, *hi*, *li*, *gi*, *u*, *gu*, or *chu*) than when they did not.

Last but not least, the frequency of occurrence of unaccusative verbs may affect learners' errors in their use. For instance, Lee et al.'s (2008) study, inspired by a usage-based approach, showed a main effect for input frequency on learners' ability to judge the grammaticality of unaccusative structures. That is, the participants found it more difficult to judge low-frequency unaccusative verbs (lower than 20 per million) than high-frequency ones (higher than 100 per million). Based on this finding, Lee et al. claim that it is likely that L2 English learners have been exposed to high-frequency unaccusative verbs more often, resulting in more solid knowledge of them.

In sum, English unaccusativity is a difficult feature to acquire, even for advanced L2 English learners, and especially for L1 Korean learners. Hence, it seems necessary to assist learners in mastering the English unaccusative structure through appropriate pedagogic intervention, and as such, English unaccusative verbs were chosen as the target construction of the present thesis. Seventeen English unaccusative verbs were identified from the two treatment texts and selected as target features. Probably due to the genre and topic of the texts, i.e., expository texts explaining and describing natural/ scientific phenomena, all of the unaccusative verbs were used in context without a conceptualizable agent (Ju, 2000). As illustrated in Table 5, all target verbs included in the texts were low frequency, and hence the participants were expected to have a limited

knowledge of those verbs. Of the 17 verbs, six verbs were non-alternating unaccusative verbs, while the rest were alternating. Each of the unaccusative verbs appeared in the texts once.

Table 5. Target English unaccusative verbs for Study 1

	Text 1				Text2		
	Unaccusative verb	Alterna -ting	Frequency (per 450 million)		Unaccusative verb	Alterna -ting	Frequency (per 450 million)
1	decompose	A	312	1	fossilize	A	11
2	subside	NA	568	2	date to	A	743
3	ascend	A	759	3	originate	A	1,022
4	accumulate	A	1,814	4	consist of	NA	2,140
5	cease	A	2,554	5	persist	NA	2,684
6	diminish	A	2,701	6	evolve	A	3,184
7	drift	NA	4,477	7	disappear	NA	7,581
8	collect	A	10,525	8	emerge	NA	9,116
9	settle	A	10,873				

Note. A = Alternating verb, NA = Non-alternating verb.

3.2.2. Pseudo-words

In addition to the English unaccusative verbs, ten lexical items were also included as target constructions in this study. The lexical targets were carefully selected from the two texts based on the following conditions: (a) the word is a noun (to control for part of speech); and (b) the word appears once (to control for frequency). Five words were selected from each text and replaced with pseudo-words that followed English orthographic and morphological rules (Pulido, 2007). When the original word was in plural form, plurality was also marked in the corresponding pseudo-word by attaching –s. Each of the pseudo-words consisted of two syllables, containing seven letters, in order to control for length.

Table 6. Target pseudo-words for Study 1

	Text 1			Text2	
	Pseudo-word	Original word		Pseudo-word	Original word
1	stragon	bottom	1	cabrons	changes
2	golands	spouts	2	fration	absence
3	phosens	discoveries	3	zenters	clues
4	klaners	parks	4	morbits	descendants
5	stovons	beaches	5	tralion	predator

4. Treatment task

The treatment task in this study was similar to what test-takers are required to do when taking the reading section of a TOEFL test, i.e., reading the text provided and answering multiple-choice comprehension questions (see Appendix B-1). In other words, a reading comprehension measure was embedded in the learning task so that the level of participants' text understanding could be simultaneously measured in tandem with task completion. The multiple choice reading comprehension items were also taken from past TOEFL tests, considering the fact that they had previously been validated by ETS (e.g., Freedle & Kostin, 1993, 1999). The reading comprehension items asked participants to identify factual/ non-factual information, make inferences, understand rhetorical purpose, recognize vocabulary meaning, determine references, simplify/ paraphrase a sentence, insert a sentence into a paragraph, and select the main ideas of a text (Educational Testing Service, 2013). As in the original TOEFL format, the texts were divided into five segments, comprised of either one or two paragraphs, and followed by reading comprehension questions relevant to each segment. There were 14 multiple-choice comprehension items for each text. One point was given to 13 items, and the last item received two points, totalling in 15 points.

In this thesis, task complexity was defined as task-induced demands imposed on learners' cognitive resources while performing a task. Drawing upon Khalifa and Weir's (2009) processing model of reading comprehension, the cognitive demands of reading tasks were manipulated in terms of the presumed depth of reading required, that is, the extent to which the task requires careful reading for successful completion. In the – complex condition, participants were asked to read and answer the comprehension questions as they normally would when working on the reading section of a TOEFL test (see Figure 7). In the + complex condition, the segments were jumbled and presented to participants in a mixed order (see Figure 8). Thus, in addition to reading the paragraphs

comprehension questions, participants in the + complex group also had to reorder the segments into a correct order to make a coherent text. Given that readers' comprehension is substantially influenced by the degree of clarity and coherence of text structure (Meyer, 1975, 1985; Meyer & Freedle, 1984; Meyer & Ray, 2011), the latter task was considered to require more careful and thorough reading than the former one. There was no time limit for task completion, as uninformed time estimations had to be collected after finishing the tasks (McClain, 1983). The total score for reading comprehension was 15 for each text (1 point for 13 items, and 2 point for 1 item).

Glossing was achieved by providing Korean definitions of the target unaccusative verbs and pseudo-word items in the margins of the texts. L1 definition glosses were chosen in order to prevent confounding variables and thereby not to make the study overly complex. For example, the efficacy of L2 glosses (i.e., synonyms or definitions in L2) can be mediated by participants' L2 proficiency, inviting another moderating variable to the research design. Also, multiple-choice glosses or fill-in tasks would have embedded a secondary task to the text-ordering task, imposing additional cognitive demands on the participants, and hence serving as a confounding variable. As mentioned earlier, Korean definitions of English unaccusative verbs often contain passive morphemes. Yet, only four target unaccusative verbs contained Korean passive morphemes, as in *naja-ci-da* (diminish), *mo-i-da* (collect), *kusongdo-i-da* (consist of) and *sara-ci-da* (disappear). The decision was made to keep using these definitions in passive voice, as they might help participants to notice and establish the mapping of target-like uses of the unaccusative verbs (in the active voice) and their passive meanings. Also, for each of the pseudo-words, a Korean definition of the original word was glossed in the margin of the text.

Directions: Read the provided passage and answer the comprehension questions.

PETROLEUM RESOURCES

Petroleum, consisting of crude oil and natural gas, has its origin from organic matter in marine sediment. Microscopic organisms settle[1] to the seafloor and accumulate[2] in marine mud. The organic matter may partially decompose[3], using up the dissolved oxygen in the sediment. As soon as the oxygen is gone, decay ceases[4] and the remaining organic matter is preserved.

Continued sedimentation – the process of deposits' settling on the sea stragon[5] – buries the organic matter and subjects it to higher temperatures and pressures, which convert the organic matter to oil and gas. As muddy sediments are pressed together, the gas and small droplets of oil may be squeezed out of the mud and may move into sandy layer nearby. Over long periods of time (millions of years), accumulations of gas and oil can collect[6] in the sandy layers. Both oil and gas are less dense than water, so they generally ascend[7] through water-saturated rock and sediment.

[1] 정착하다, 자리잡다
[2] 늘어나다
[3] 부패하다
[4] 멎다, 정지하다
[5] 바닥
[6] 모이다
[7] 오르다, 뜨다

1. The word "crude" in the passage is closest in meaning to
   (a) unseen
   (b) unprocessed
   (c) uncovered
   (d) unnatural

Figure 7. Sample task layout of – complex condition for Study 1

Directions: Read the paragraphs A – E and answer the comprehension questions. Also, rearrange them to make a coherent text.

Rearrange the paragraphs A – E in a correct order:

_____ – _____ – _____ – _____ – _____

PETROLEUM RESOURCES

A. Of course, there is far more oil underground than can be recovered. It may be in a pool too small or too far from a potential market to justify the expense of drilling. Some oil is located under regions where drilling is forbidden, such as national klaners[1] or other public lands. Even given the best extraction techniques, only about 30 to 40 percent of the oil in a given pool can be brought to the surface. The rest is far too difficult to extract and has to be left underground.

[1] 공원

1. According to the paragraph above, the decision to drill for oil depends on all of the following factors EXCEPT
   (a) permission to access the area where oil has been found
   (b) the availability of sufficient quantities of oil in a pool
   (c) the location of the market in relation to the drilling site
   (d) the political situation in the region where drilling would occur

Figure 8. Sample task layout of + complex condition for Study 1

5. Assessment tasks

In this study, knowledge of the target constructions was operationalised as (a) the ability to recognize the grammaticality of English unaccusative verbs and (b) the ability to recognize the form and meaning of the target lexical items. Learning of the target unaccusative verbs was measured with an untimed grammaticality judgment test in a written mode (henceforth, GJT). An untimed GJT was employed in order to tap into participants' implicit as well as explicit knowledge of English unaccusative verbs. It was also expected that, when supplemented by reaction time recordings, confidence ratings and subjective source attributions, the test would enable fuller and more valid assessment of the nature of the acquired knowledge by the participants. In addition, the GJT was constructed in the written mode, considering the fact that the participants performed reading tasks that entailed processing of the target constructions in the form of textual input.

Learning of the pseudo-words was assessed via multiple-choice form and meaning recognition tests. Given that the participants were exposed to each of the target pseudo-words only once while performing the reading tasks, form and meaning recognition tests were considered as appropriate to measure the participants' knowledge of the items (Laufer & Goldstein, 2004).

5.1. Grammaticality judgment test

As aforementioned, participants' knowledge of English unaccusative verbs was measured with an untimed GJT accompanied by reaction time recordings, binary confidence ratings, and subjective source attributions. In previous research, reaction time data has proved to be a valid supplement to GJTs in revealing the source and solidity of knowledge that underlies learners' responses (e.g., Bley-Vroman & Masterson, 1989; Jiang, 2011; Juffs, 2001; L. White & Juffs, 1998). That is, it has been argued that ungrammatical sentences may take longer to be judged because there is no

structural representation of them in the learner's internal grammar. As a result, the parser may try different analyses before giving up and labelling a sentence as ungrammatical (Juffs, 2001). Binary confidence ratings (e.g., Kunimoto, Miller, & Pashler, 2001; Rebuschat & Williams, 2012) and the subjective attribution of source knowledge (e.g., Dienes & Scott, 2005) have been suggested as offering estimates of the conscious or unconscious status of participants' knowledge.

The GJT contained 80 sentences in total, including (a) 34 sentences for the target unaccusative verbs, (b) 16 for novel unaccusative verbs and (c) 30 distractors (see Appendix B-1). First, for each of the 17 target unaccusative verbs, one grammatical and one ungrammatical passive sentence were created, resulting in 34 sentences.

Also, in order to explore if acquired knowledge was transferred to other unaccusative verbs, eight additional verbs were selected from the list of the 2,000 most frequently used English words, by consulting *Compleat Lexical Tutor* version 6.2. The selection was guided by previous studies that reported English unaccusative verbs Korean learners typically have difficulty in acquiring (Hwang, 1999, 2001; No & Chung, 2006; Shin, 2011). As shown in Table 7, among the eight verbs, four were non-alternating unaccusative verbs and four were alternating ones. Additionally, Korean definitions of four verbs contained passive morphemes, whereas those of the other four did not. For these eight verbs, eight grammatical and eight ungrammatical sentences were produced, a total of 16 sentences. Care was taken to control the number of syllables, syntactic complexity, semantic plausibility, vocabulary familiarity and the position of unaccusative verbs for each of the 25 pairs of unaccusative sentences (see Bley-Vroman & Masterson, 1989).

Lastly, 30 sentences, 15 grammatical and 15 ungrammatical, were included as distractors. The grammatical rules the distractors drew on included gerunds, to-infinitives, subjective moods, comparatives, participial adjectives, reflexives, relative

pronouns, inversion and prepositions, which cover the topics generally dealt with in English grammar lessons in Korea (No & Chung, 2006). Five native speakers took part in a pilot test to ensure there were no grammatically ambiguous or vague sentences included. Across the pretest, posttest and delayed posttest, the same 80 sentences were randomly presented to participants.

Table 7. Additional unaccusative verbs

	Non-alternating				Alternating		
	Unaccusative verb	Frequency (per 450 million)	Korean definition		Unaccusative verb	Frequency (per 450 million)	Korean definition
1	occur	18,880	*irona-da*	1	burn	11,690	*ta-da*
2	remain	37,993	*nam-da*	2	stop	86,198	*momchu-da*
3	appear	36,739	*bo-**i**-da*	3	break	72,852	*ke-**ci**-da*
4	fall	67,590	*toro-**ci**-da*	4	change	138,913	*baku-**i**-da*

Note. *i* and *ci* are Korean passive morphemes.

The GJT for this study was constructed using E-Prime 2.0 in order to allow for reaction time analysis. Eighty sentences were sequentially presented on a computer screen and participants were asked to press the "*z*" key if a sentence seemed grammatical and the "*m*" key if it seemed ungrammatical. These particular keys were chosen by considering their placement on a QWERTY keyboard, which is the normal layout in Korea. Each sentence remained on the screen until a decision on the well-formedness of the sentence was made, and participants were instructed to make their decisions as fast as they could. In order to measure reaction times, a timer started at the onset of each sentence and stopped when a response was given. After a response was given, participants were asked to make a binary decision depending on their level of confidence in the response (*high* vs *low confidence*). After a confidence rating, they were asked to select the source of their decision from four options: *guess*, *intuition*, *rules* and *memory*. Between each set of grammaticality judgment item, confidence rating and source attribution task, a fixation cross appeared in the centre of the screen for 500 milliseconds to signal an upcoming sentence (see Figure 9). The total score was

50 (34 for target verbs and 16 for novel verbs), and the test took approximately 10–12 minutes to complete.

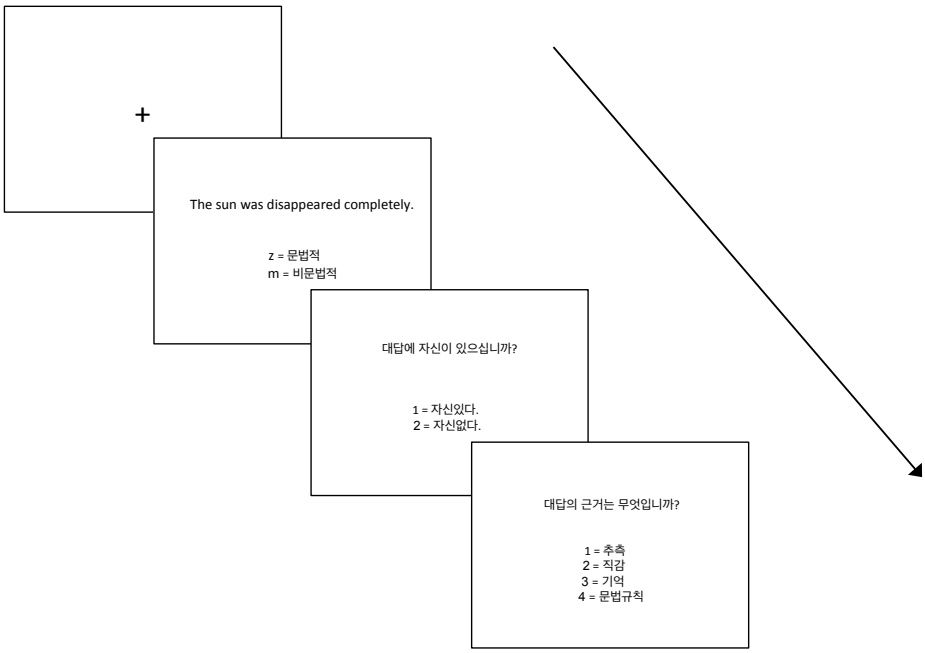

Figure 9. Example of slides used in the grammaticality judgment test

5.2. Vocabulary form recognition test

In order to measure if task complexity and glossing affected form recognition of the target pseudo-words, 20 items were constructed using E-Prime 2.0 (see Appendix C-1). The items were modelled after a similar task used in Leeser's (2007) study. Participants were asked to press either "*z* (yes)" or "*m* (no)", depending on whether they remembered seeing the word in the texts. Ten items were target pseudo-words, whereas the other ten were distractors that were constructed drawing on the pseudo-words in Godfroid et al. (2013). Each of the distractors contained two syllables and seven letters as target pseudo-words. The 20 items were randomized and presented on a computer screen, and participants were asked to choose an answer for each item, followed by a binary decision task asking the level of confidence in their response (*high* vs *low confidence*). Again, participants were instructed to answer each item as fast as they could, and response latency was recorded in milliseconds for each item in order to infer

meaningful information as to the source and robustness of learning (Jiang, 2011). Between each set, comprising a form recognition item and a binary confidence rating, a fixation cross appeared on the screen for 500 milliseconds to signal the next item (Figure 10). The total score for this test was 10 (1 for each correct and 0 for each incorrect item), and the test took approximately 2–3 minutes.

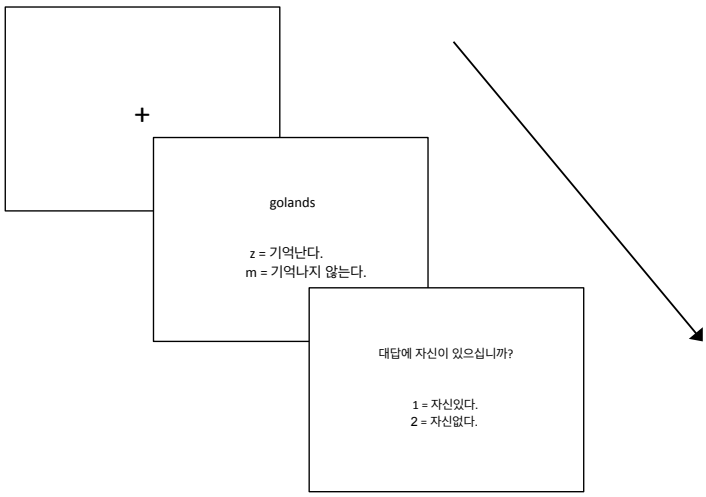

Figure 10. Example of slides used in the word form recognition test

5.3. Vocabulary meaning recognition test

In addition to the 20 form recognition items, 20 additional multiple-choice items, modelled after Martinez-Fernández's (2010) meaning recognition test, asked participants to select a correct Korean definition of a given target word from three options (see Appendix D-1). Among these, ten items were target words while the other ten were the distractors used in the form recognition test. The multiple-choice options included the gloss used for the target word, two glosses used for other target words and "*I don't know*", which was to prevent participants guessing. As in the form recognition test, the items were randomized and presented on a computer screen, immediately followed by binary confidence ratings. The procedure was the same as in the word form recognition test (see Figure 11). The total score was 10 (1 for each correct and 0 for each incorrect item), and the test took approximately 3 minutes.

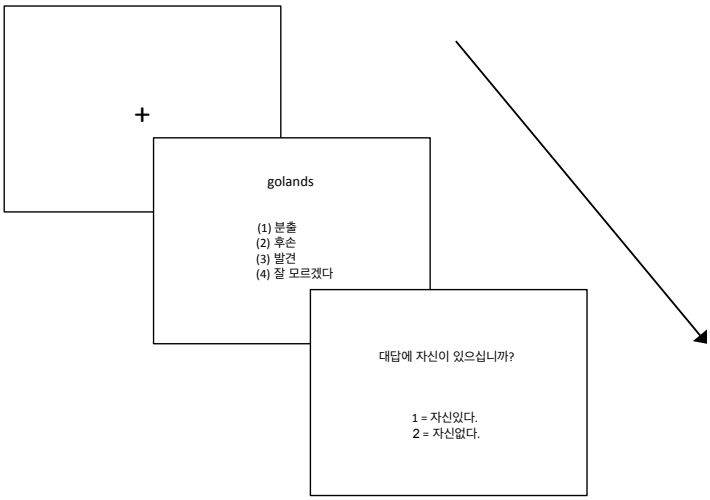

Figure 11. Example of slides used in the word meaning recognition test

6. Working memory measures

In this thesis, a forward digit span test and a nonword repetition test were used to measure participants' phonological short-term memory. They both required immediate recall of a series of unrelated digits or nonwords in the order of presentation, measuring the participants' ability to retain phonological information in short-term memory. In addition, a backward digit span test and an automated operation span test were used to measure participants' complex working memory. These tests necessitated intentional manipulation of incoming information while retaining it in short-term memory, which was designed to assess the participants' executive control function as well as phonological short-term memory.

6.1. Forward digit span test (DS)

In the forward digit span test, adopted from Brunfaut and Révész (2015), participants were provided with sequences of unrelated digits that were presented in an automated PowerPoint slide show. Each digit stayed in a slide for 1 second, and set sizes ranged from 3 to 11 digits (2 sets for each length, 18 sets in total) presented in an increasing order. Digits were repeated across sets but not within sets, and all of them were used approximately equally in the test. Participants were provided with a response

sheet, modelled after Kane, Hambrick, Tubolski, Wilhelm, Payne, and Engle (2004), which contained 18 rows of nine blank spaces, with each row representing one set. Participants were instructed to recall the digits from each set in the response sheet, with one digit in each blank. Ten seconds were allowed for recalling each set. The maximum set size correctly recalled once was the digit span score. Cronbach's alpha for the test was .76. The test took about 7–8 minutes.

6.2. Nonword repetition test (NWS)

For this study, a nonword repetition test was developed in Korean. More specifically, nonsense words that conformed to the phonotactic rules of Korean were created and then presented to participants in an automated PowerPoint slide show. The test stimuli consisted of 32 nonwords, each containing 4 to 11 syllables (4 sets for each syllable length). Each nonword was aurally delivered to participants in a random order, and ten seconds were allowed for oral recall. Each of the nonword recalls was scored either correct or incorrect, and the maximum number of syllables that participants correctly recalled at least twice for each syllable length was the score for this test. The test was piloted on seven Korean graduate students to determine appropriate syllable lengths and the reliability of the test. They were also asked to rate the *wordlikeness* of each nonword on a 5-point Likert scale from 1 (very likely to pass for a real Korean word) to 5 (very unlikely to pass for a real Korean word). This process was to ensure that the nonword stimuli of the test were low in wordlikeness so that participants would be less likely to retrieve similar phonological structures from their long-term memory and have to depend on short-term phonological representation to mediate nonword repetition (Gathercole, 1995). The mean value of wordlikeness was 2.23 (*SD* = .74). Seven nonwords that were rated relatively highly for wordlikeness (1 *SD* above from the mean) were replaced with other less-wordlike nonwords. Cronbach's alpha for this

test was .73. The test was administered individually, and took about 9–10 minutes to complete.

6.3. Backward digit span test (BDS)

The design, structure and procedure of the backward digit span test (Brunfaut & Révész, 2015) were the same as for the forward digit span test, except for the fact that participants were instructed to recall the digits in reverse order. The maximum set size correctly recalled once was the backward digit span score. The test took about 7–8 minutes, and the Cronbach's alpha was .81.

6.4. Automated operation span test (OSPAN)

An operation span test, created by Turner and Engle (1989), requires participants to solve a series of math problems while remembering a set of unrelated letters or words. As opposed to language-specific aspects of the reading span task (e.g., Daneman & Carpenter, 1980; Waters & Caplan, 1996), the operation span task taps into general complex working memory capacity. The source file of the automated operation span test, constructed for E-Prime 2.0.10.242, was obtained from the *Attention and Working Memory Lab* at Georgia Tech (Unsworth, Heitz, Schrock, & Engle, 2005; Redick, Broadway, Meier, Kuriakose, Unsworth, Kane, & Engle, 2012) and regenerated for E-Prime 2.0.10.353. The automated version of the operation span test allows participants to complete the test independently by clicking a mouse button. The test began with two practice sessions to familiarize participants with the math operation and letter recall tasks and to calculate individual differences in the time required to solve the math problems (see Figure 12). The time taken to solve the math problems (plus 2.5 SD, determined after extensive piloting; Unsworth et al., 2005) was used as the time limit for each math problem session for that individual. In order to guarantee that participants engaged in a trade-off between storage (remembering letter strings) and processing

(solving math problems), an 85% accuracy criterion for the math operation was required. The real test session consisted of three sets, with set sizes ranging from 3 to 7 (75 letters and 75 math problems in total). The order of set size was random. Following Unsworth et al. (2005), the total number of correct letter recalls was used as the OSPAN index. The test took approximately 20 minutes to complete. According to Unsworth et al. (2005), Cronbach's alpha for the test was .78.

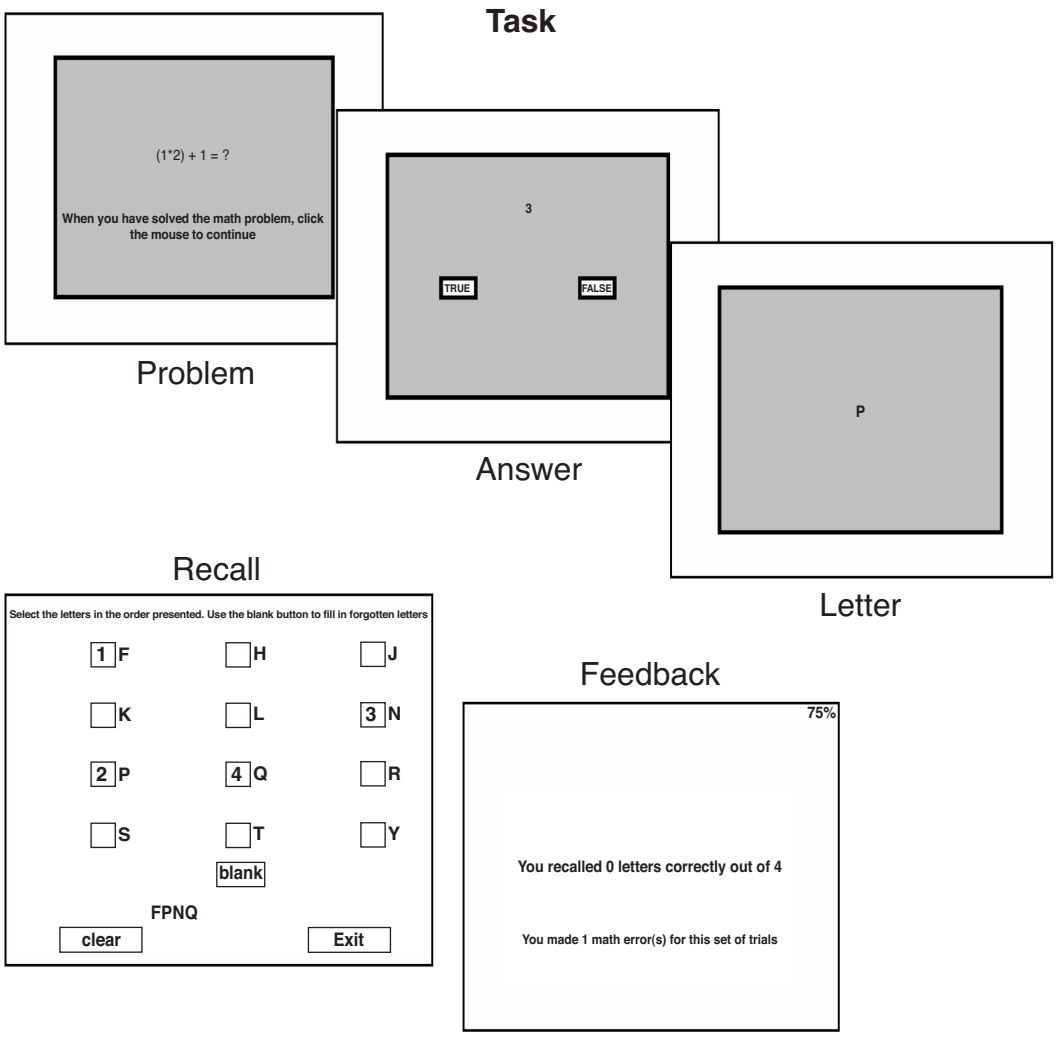

Figure 12. Example of slides used in the OSPAN test (*Source*: Adopted from Unsworth et al., 2005, p. 500)

7. Questionnaires

Participants were asked to complete a background questionnaire, two post-reading questionnaires and an exit questionnaire (see Appendix E). The aim of the background questionnaire was to collect information about participants' demographics and English

learning experiences. The post-reading questionnaires asked participants to provide their retrospective subjective time estimation taken to complete the given reading task, perceived level of mental effort invested in task completion, and familiarity with the topic of the reading text. The rationale for using the retrospective subjective time estimation was that (a) it is relatively easy to do and (b) it can be used as an additional source for estimating the cognitive/ mental demand imposed on learners (Baralt, 2013; Block, Hancock, & Zakay, 2010; Fink & Neubauer, 2001; Thomas & Weaver, 1975). As the time estimation task was conducted under a retrospective paradigm (Fink & Neubauer, 2001), participants were unaware of the upcoming duration judgment task until it had to be done. As such, subjective time estimations were only collected after the first treatment session. Unlike the prospective paradigm (informed time estimation), wherein estimated-to-real duration ratio decreases with increasing cognitive load, it was expected that the duration ratio would increase after performing cognitively more demanding tasks under the retrospective paradigm. The perception questionnaire items tapped into various aspects of difficulty, such as stress, perceived ability to complete the task, interest in the task, confidence in task performance, mental effort invested and motivation to complete the task. Two items were constructed to measure each of the sub-constructs. Finally, an exit questionnaire asked participants to make comments on their experience of task performance, e.g., if they learned any English linguistic features from this study or if they studied any English lexical or grammatical items outside of this study. All questionnaires were administered in Korean.

II. Procedure

Prior to the outset of data collection, ethical approval was obtained from the UCL Institute of Education. As shown in Figure 6, data were collected over four weeks. In the first session, all participants signed a consent form after reading an information

sheet that explained (a) the right to withdraw from the research at any time, (b) the overall structure and procedure of the research, (c) potential inconveniences and benefits of participation and (d) measures taken to maintain their privacy and confidentiality (see Appendix A-1). Next, they took the pretest, a background questionnaire and an L2 proficiency test (CPE) in the first session. One week later, participants took part in two treatment sessions on separate days. In each of the treatment sessions, they performed a reading task (texts were counter-balanced in each condition) and answered a post-reading questionnaire immediately after task completion. In the third session, they completed an additional immediate posttest after finishing the second treatment task. Two weeks later, participants completed a delayed posttest and were subjected to working memory tests. Each session took approximately 45 minutes to an hour. The experimental sessions were conducted in a computer laboratory at a university in Korea. In the course of data collection, the participants did not receive any instruction on English unaccusative verbs outside of the study.

III. Analysis

1. Statistical analyses

SPSS 22.0 (Statistical Package for the Social Sciences) for Mac was used to examine the reliability of the tests as well as compute descriptive and correlational statistics for the data. More specifically, the reliability of the different tests was determined using Cronbach's alpha, and interrelationships between the various test scores were computed using Pearson's coefficient. The level of significance for this study was set at an alpha level of $p < .05$. Mixed-effects models were constructed to examine mean differences among the groups in terms of CPE scores, pretest GJT scores and ratings of topic familiarity and perceived task difficulty. Also, mixed-effects modelling was used to explore the effects of the independent variables (i.e., task

complexity and glossing) and moderating factors (i.e., working memory capacity measures) on the dependent variables (i.e., reading comprehension scores, GJT scores and vocabulary recognition scores). In order to do this, the statistical program R version 3.3.0 was used (R Development Core Team, 2016). The rationale for doing mixed-effects modelling and the detailed procedures of this study is explained in the following section.

2. Mixed-effects modelling in R

Mixed-effects regression analyses have received increasing interest among SLA researchers, because they offer several advantages over standard regression analyses. One particular strength of mixed-effects modelling is that it can account for the potential idiosyncrasies nested in participants and items (Baayen, 2008; Cunnings, 2012; Linck & Cunnings 2015; Rogers, 2016; Winter, 2013), and thus allow researchers to make a "simultaneous generalization of the results on new items and new participants" (Gagné & Spalding, 2009, p. 25). In addition, mixed-effects modelling can be used for either interval-scale or categorical-scale data by producing linear or logit models, respectively, and it can also manage missing values and imbalanced research designs (Sonbul & Schmitt, 2013). Considering that this study included participant- and item-related factors as well as both interval-scale (e.g., reaction time data) and categorical-scale measures (e.g., correct vs incorrect responses), and that outliers were removed resulting in an imbalanced research design, mixed-effects modelling was considered a robust and appropriate method for data analysis.

Prior to constructing the models, test scores and reaction times that were outside of +/–1.5×IQR (the third quartile – the first quartile) were specified as outliers. Additionally, for each participant, reaction times given in response to incorrectly answered items and values that were 2 *SD*s longer or shorter than the mean for the participant were removed from the data (de Zeeuw, Verhoeven, & Schreuder, 2012;

Jiang, 2007). As a result, 30.96% of the reaction time data were removed from the data set. Next, for interval-scale data, residual plots were drawn to check linearity and homoscedasticity. As shown in Figure 13, residual plots of reaction time data revealed heteroscedasticity (i.e., funnel shaped distribution of residuals), so a logarithmic transformation was performed to fix the problem (Baayen, 2008).

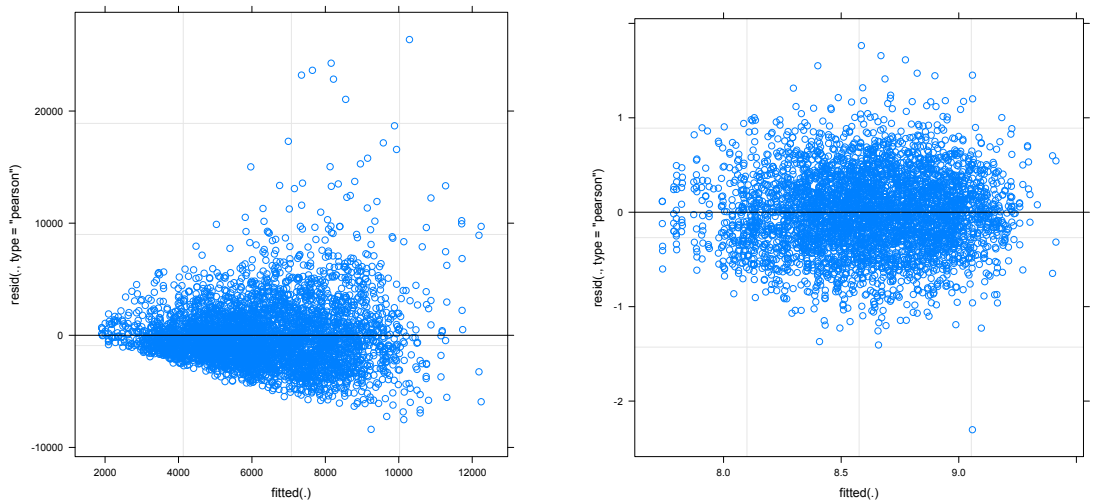

Figure 13. Residual plots of RT data before and after logarithmic transformation

The data were then analysed by constructing various mixed-effects models with the package *lme4* (Bates, Maechler, & Bolker, 2012). The fixed effects in all the models were the two independent variables, Complexity and Glossing. As recommended by Linck and Cunnings (2015), contrast coding was implemented in order to reduce the likelihood of correlation between fixed effects (e.g., -.5 – complex, .5 + complex; -.5 unglossed, .5 glossed). If the fixed effects correlate, which means they tap into overlapping constructs, the significance of each fixed effect becomes difficult to interpret as they may share each other's explanatory power (Winter, 2013). Thus, collinearity, i.e., the degree to which fixed effects correlate with each other, should be low enough to ensure the stability of the model. As random effects, the models included intercepts for Subjects and Items, as well as by-Subject and by-Item random slopes for the fixed effects (Barr, Levy, Scheepers, & Tily, 2013). For GJT data, Time was put as

an additional fixed effect in order to explore changes in the data over repeated measurements and Grammaticality to compare participants' performance between grammatical versus ungrammatical sentences. Finally, working memory capacity indices were also added as fixed effects to the models, one by one, when exploring the moderating role of working memory capacity.

For interval-scale data, such as reading comprehension scores and reaction times, linear mixed-effects models were constructed using the *lmer* function. For categorical-scale data, such as correct versus incorrect responses to the GJT and vocabulary recognition tests, logit models were produced using the *glmer* (*generalized* linear mixed model) function with the argument *family=binomial* (Linck & Cunnings, 2015). The modelling started by constructing a null model that contained only the random intercept of Subject and Item (e.g., *rtnull <- lmer(log(rt) ~ (1|Subject) + (1|Item), data = rtdata)*). Next, each of the fixed effects was entered into the null model step-wisely (e.g., *rtcom <- lmer(log(rt) ~ complexity + (complexity|Subject) + (complexity|Item), data = rtdata)*) and tested to see whether the inclusion of the fixed effect significantly improved the fit of the model. As part of this step, likelihood ratio tests were computed using the χ^2 statistic (e.g., *anova(rtnull, rtcom)*).

After identifying the fixed effects that improved the null models significantly, maximal random effects structures were produced following Barr et al. (2013). As mixed-effects models with a maximal random structure can be overly complex, with multiple random slope parameters, models often fail to converge. In this case, the first step was to run the model with the "*bobyqa*" optimizer (Powell, 2009), which is for stabilizing a model. If this did not resolve the convergence issue, the random effect parameters that accounted for the least variance were removed in a step-wise fashion until convergence was achieved (Blom, Paradis, & Sorenson Duncan, 2012; Cunnings & Sturt, 2014). Next, in order to identify the best fitting model, each fixed effect was

removed one by one from the full model and tested against the full model. If the removal of a fixed effect improved the model fit, the fixed effect was excluded.

Before running the best fitting model generated by following the aforementioned procedure, any influential data points were identified using the R package *influence.ME* (Nieuwenhuis, Pelzer, & te Grotenhuis, 2012). This step was taken, in addition to the removal of outliers at the individual level, in order to double-check whether there was any remaining value that might drastically affect the interpretation of the results. *influence.ME* calculated *DFBETAS*, Cook's distance and a test for changing levels of significance, while accounting for the nesting structure of the data. *DFBETAS* estimates the level of influence that observations have on each of the fixed effects (Fox, 2002), and the cut-off value was set at $2/\sqrt{n}$ where n refers to the number of groups in the grouping factor (Belsley, Kuh, & Welsch, 1980). Cook's distance provides a summary value for all fixed effects simultaneously. The cut-off value for Cook's distance was set at $4/n$ where n, again, refers to the number of groups in the grouping factor (Van der Meer, te Grotenhuis, & Pelzer, 2010). Also, using the *sigtest* function, it was tested whether the exclusion of each single case changed the statistical significance of the fixed effects in the models. This procedure revealed no influential data point for any of the models.

While the results of logit models provide *p*-values for *z* statistics, linear model summaries provide *t* statistics without *p*-values. Hence, absolute *t*-values above 2.0 were set as a criterion for testing the significance of the models (Gelman & Hill, 2007). Effect sizes for the linear mixed-effects models were computed using *r.squaredGLMM* function from the package *MuMln* (Barton, 2015), whereas that of the logit mixed-effects models was calculated with *C* index of the concordance using the *somer2* function in the *Hmisc* package (Harrell & Dupont, 2015). Following Plonsky and Oswald (2014), R^2 values of .06, .16 and .36 were interpreted as small, medium and

large, respectively. A *C*-index of .70 was considered a moderate fit, .80 good and .90 and above excellent for the data (Baayen, 2008; Gries, 2013; Rogers, 2016). For *t*-tests, Cohen's *d* was calculated to examine effect sizes. As suggested by Plonsky and Oswald, the benchmarks were .40 for small, .70 for medium and 1.00 for large effect sizes for independent-sample *t*-tests, and .60 for small, 1.00 for medium and 1.50 for large effect sizes for paired-sample *t*-tests. Collinearity statistics for the independent variables, i.e., Complexity and Glossing, were calculated using *collin.fnc* function in the *languageR* package (Baayen, 2008; Belsley et al., 1980). Following Baayen, condition numbers between 0 and 6 were regarded as no collinearity, around 15 as medium, and 30 or above as potentially harmful collinearity.

VI. Results

1. Preliminary analysis

Before answering the research questions, some preliminary steps were taken to ensure the reliability of the instruments and the validity of the results. The following methodological concerns were taken into consideration: reliability of the tests, participants' prior knowledge of the target items, potential effects of topic familiarity on reading comprehension scores, and validity of task complexity manipulation.

1.1. Test reliability

In order to check the consistency and stability of the instruments, reliability coefficients for the proficiency test, reading comprehension tests, grammaticality judgment tests and vocabulary recognition tests were computed using Cronbach's alpha. As summarized in Table 8, the values for Cronbach's alpha were found to be high for the proficiency and grammaticality judgment tests, but low for the reading comprehension and vocabulary meaning recognition tests. Also, the mean score for the vocabulary meaning recognition test was very low, presumably contributing to the low

reliability coefficient. In addition, the mean reading comprehension scores imply a ceiling effect, which could have contributed to the low internal consistency reliability of the reading comprehension tests.

Table 8. Descriptive statistics for test scores

	N	Mean	*SD*	*Cronbach's alpha*
CPE test	52	14.29	4.94	.70
Reading comprehension (Text 1)	52	11.04	2.22	.47
Reading comprehension (Text 2)	52	12.85	1.51	.37
Grammaticality judgment (Target items)	52	58.52	11.77	.80
Grammaticality judgment (Novel items)	52	31.92	6.07	.67
Vocabulary recognition (Form)	52	11.64	3.65	.59
Vocabulary recognition (Meaning)	52	5.64	2.96	.45

Note. Maximum score for: CPE test = 45, reading comprehension = 15, grammaticality judgment (target) = 102, grammaticality judgment (novel) = 48, vocabulary form recognition = 20, vocabulary meaning recognition = 20.

1.2. Equivalence among groups

To check the equivalence of English proficiency level among the groups, a mixed-effects model was constructed, with CPE scores as the dependent variable, Group as a fixed effect, and Subject and Item as random effects. When compared with a null model that contained only random effects, the results showed that the inclusion of Group as a fixed effect did not make a significant difference to the null model, $\chi^2(1) = .29$, $p = .59$, $R^2 < .01$. In other words, the groups did not significantly differ from each other in terms of their English proficiency (for descriptive statistics, see Table 9).

Table 9. Descriptive statistics for proficiency test

	Proficiency test			
Group	*N*	Mean	*SD*	*SE*
[– C, – G]	13	14.39	5.30	1.93
[– C, + G]	13	13.54	4.81	1.27
[+ C, – G]	13	14.23	4.38	1.39
[+ C, + G]	13	15.00	5.66	1.47
Total	52	14.29	4.94	.68

Note. Maximum score = 45.

Next, in order to test whether the four groups started out at a developmentally parallel stage, another set of likelihood ratio tests were conducted on pretest GJT scores, comparing null models only with random effects and models additionally containing

Group as a fixed effect (for descriptive statistics, see Table 15). The results indicated that Group did not improve the null models to a significant degree, Target verbs: $\chi^2(1) = .25$, $p = .62$, $R^2 < .01$; Novel verbs: $\chi^2(1) = 1.14$, $p = .29$, $R^2 < .01$. In sum, the results showed that, at the time of the pretest, there were no significant differences among the groups in their ability to judge the grammaticality of English unaccusative sentences.

1.3. Effects of topic familiarity

To assess the potential impact of topic knowledge on comprehension of the treatment texts, participants' familiarity with the two topics was measured using post-reading questionnaire items (i.e., Item 6: *I thought this reading topic was familiar*, Item 13: *I had some background knowledge about the reading topic*). Descriptive statistics are presented in Table 10. The responses to the two items significantly correlated with each other, Text 1: $r(52) = .68$, $p < .01$, Text 2: $r(52) = .56$, $p < .01$, suggesting that the items assessed overlapping constructs.

Table 10. Descriptive statistics for topic familiarity by item

		Topic familiarity					
		Text 1			Text 2		
Item	*N*	Mean	*SD*	*SE*	Mean	*SD*	*SE*
# 6	52	3.60	.24	1.75	3.10	.24	1.76
#13	52	3.48	.23	1.69	2.46	.19	1.34
Total	52	7.08	.44	3.16	5.55	.38	2.75

Note. Maximum value for each item = 7.

In order to examine the effects of topic familiarity on reading comprehension scores, likelihood ratio tests were conducted, comparing a null model with Subject and Item as random effects and models additionally including topic familiarity as a fixed effect. The dependent variable was reading comprehension scores for Text 1 and Text 2. The results showed that adding topic familiarity did not make significant improvement to the null models, Text 1: $\chi^2(1) = .01$, $p = .91$, $R^2 < .01$, Text 2: $\chi^2(1) = 2.25$, $p = .13$, $R^2 < .01$. In short, the participants' topic familiarity with the texts did not affect their scores on reading comprehension items.

1.4. Validation of task complexity manipulation

To validate the operationalisation of task complexity, all participants were asked to judge the time taken to complete each task immediately after task completion. As mentioned earlier, only time estimations made after completing the first task were analysed. In order to examine whether subjective time estimations differed as a function of task manipulation (Block et al., 2008), estimated-to-target duration ratios were calculated by dividing the estimated time by the real time taken to complete a given task. Hence, a duration judgment ratio higher than 1 indicated that participants overestimated the time taken to complete a task, as compared to the actual time spent on the task. In the retrospective time estimation paradigm, the duration judgment ratio is expected to increase with greater cognitive load.

Table 11. Descriptive statistics for duration judgment ratio

Condition	*N*	Text 1 Mean (*SD*)	Text 2 Mean (*SD*)	Total Mean (*SD*)
– Complex	13	1.03 (*.16*)	.95 (*.11*)	.99 (*.15*)
+ Complex	13	1.16 (*.17*)	1.19 (*.29*)	1.17 (*.24*)

As shown in Table 11, the duration judgment ratios of both Text 1 and Text 2 in the + complex conditions were on average larger than those in the – complex conditions. The results from independent-samples *t*-tests on duration judgment ratios across + and – complex conditions also revealed significant effects of task complexity for both Text 1 and Text 2, Text 1: $t(50) = 2.86$, $p = .01$, 95% CI [.04, .22]; Text 2: $t(50) = 3.85$, $p < .01$, 95% CI [.11, .36]. Cohen's *d*s were .79 and 1.09, respectively, which were considered medium and large effect sizes. In other words, duration judgment ratios in the + complex conditions were significantly greater than those in the – complex conditions, suggesting that + complex tasks induced heavier cognitive load on the participants compared to– complex tasks.

To infer the effects of task complexity on the amount of mental effort imposed on the participants, three questionnaire items were included in the post-task questionnaires

(Item 1: *I thought this task was difficult*, Item 7: *I invested a large amount of mental effort to complete this task*, Item 14: *I thought this task was demanding*). Cronbach's alpha for the three items was .63 for Text 1 and .75 for Text 2. Descriptive statistics for the responses to the three items are presented in Table 12.

Table 12. Descriptive statistics for perceived task difficulty by item

			Reported mental effort					
			Text 1			Text 2		
Item	Condition	*N*	Mean	*SD*	*SE*	Mean	*SD*	*SE*
# 1	– Complex	26	4.23	1.03	.20	3.77	1.03	.20
	+ Complex	26	4.65	.80	.16	4.46	1.07	.21
# 7	– Complex	26	5.12	1.11	.22	4.69	1.12	.22
	+ Complex	26	4.85	1.19	.23	4.42	1.42	.28
# 14	– Complex	26	4.00	1.17	.23	3.62	1.10	.22
	+ Complex	26	4.46	1.42	.28	4.19	1.27	.25
Total	– Complex	26	13.35	2.50	.49	12.08	2.79	.55
	+ Complex	26	13.96	2.71	.53	13.08	3.02	.59

Note. Maximum value for each item = 7.

In order to examine if there were significant differences between the + and the – complex conditions in participants' ratings of perceived task difficulty, likelihood ratio tests were conducted on the responses to the reported mental efforts. Null models included only random effects (i.e., Subject and Item), whereas increased models contained Complexity as a fixed effect. Significance was found for Text 2: $\chi^2(1) = 88.99$, $p < .01$, $R^2 = .06$, but not for Text 1: $\chi^2(1) = 2.75$, $p = .10$, $R^2 = .12$. Summaries of the increased mixed effects model for Text 2 revealed that participants in the + complex conditions rated the amount of mental effort significantly greater than those in the – complex conditions, Text 2: *Estimate* = .96, $t = 4.63$. The effect size of the model was evaluated as small.

2. Effects of task complexity and glossing on L2 reading

This section presents results that address RQ (1) and RQ (3). For the convenience of the reader, those RQs are repeated here:

RQ (1) To what extent do the cognitive demands of second language reading tasks affect reading comprehension?

RQ (3) To what extent does glossing affect second language reading comprehension?

First, descriptive statistics for the reading comprehension scores of each group are displayed in Table 13. Reading comprehension scores for Text 2 were on average higher than those for Text 1.

Table 13. Descriptive statistics for reading comprehension scores

		Text 1			Text 2		
Group	*N*	Mean	*SD*	*SE*	Mean	*SD*	*SE*
[– C, – G]	13	11.15	1.35	.37	13.23	1.42	.40
[– C, + G]	13	10.85	2.91	.81	12.00	2.00	.56
[+ C, – G]	13	11.31	2.53	.70	13.08	1.44	.40
[+ C, + G]	13	10.85	2.04	.56	13.08	.76	.21
Total	52	11.04	2.22	.31	12.85	1.85	1.51

Note. Maximum score = 15.

To examine whether task complexity and glossing had a significant impact on L2 reading comprehension scores, linear mixed-effects models were constructed with R. Null models contained random effects (i.e., Subject and Item) only, and each of the fixed effects (i.e., Complexity and Glossing) was entered one by one and compared against the null models with likelihood ratio tests using χ^2 statistics. As summarized in Table 14, neither task complexity nor glossing had significant effects on reading comprehension scores.

Table 14. Summary of likelihood ratio tests for predictors on reading comprehension scores

		χ^2	*df*	*p*	R^2
Text 1	Complexity	.01	1	.90	.00
	Glossing	.39	1	.53	.00
Text 2	Complexity	.40	1	.53	.00
	Glossing	.07	1	.79	.00

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

Lastly, in order to see if interaction of Complexity and Glossing had significant influence on reading comprehension scores, another likelihood ratio tests were carried out, comparing a reduced model that included Complexity and Glossing as separate fixed effects (e.g., Complexity + Glossing) and the one that contained interaction between the two fixed effects (e.g., Complexity * Glossinig). The results showed that

reading comprehension scores were not influenced by interaction between the two fixed effects, $\chi^2(1) = .06$, $p = .81$, $R^2 < .01$.

3. Effects of task complexity and glossing on L2 learning

This section presents the results for the effects of task complexity and glossing on development in the knowledge of target constructions. More specifically, the following research questions are addressed:

RQ (2) To what extent do the cognitive demands of second language reading tasks affect development in the knowledge of target language constructions?

RQ (4) To what extent does glossing affect development in the knowledge of target language constructions?

Results for the effects of task demands and glossing on participants' knowledge of unaccusative verbs are summarized first, followed by the results for vocabulary form and meaning recognition scores.

3.1. Effects of task complexity and glossing on learning unaccusative verbs

Table 15 presents descriptive statistics for the GJT scores by group. It appears that mean gain scores were, overall, higher in the delayed posttest than in the immediate posttest.

Table 15. Descriptive statistics for gains in the grammaticality judgment test

Group	Test	*N*	Target items (*n* = 34)			Novel items (*n* = 16)		
			Mean	Mean gain	*SD*	Mean	Mean gain	*SD*
[– C, – G]	Pretest	13	18.23		4.25	10.92		2.10
	Immediate posttest	13	20.00	1.62	5.24	11.08	.31	2.06
	Delayed posttest	13	19.39	1.00	5.01	11.39	.62	2.10
[– C, + G]	Pretest	13	17.62		4.59	10.08		2.99
	Immediate posttest	13	19.15	1.54	5.03	9.69	-.39	3.28
	Delayed posttest	13	21.69	4.08	4.68	10.92	.85	2.40
[+ C, – G]	Pretest	13	17.31		3.23	9.69		2.69
	Immediate posttest	13	18.46	1.54	3.50	10.85	.85	2.58
	Delayed posttest	13	20.23	2.85	4.19	11.23	1.54	2.09
[+ C, + G]	Pretest	13	17.39		3.62	10.39		1.81
	Immediate posttest	13	21.08	3.69	4.27	10.69	.31	2.69
	Delayed posttest	13	23.23	5.85	3.44	11.23	.85	2.32

Note. Maximum score for: target GJT items = 34, novel GJT items = 16.

To examine whether task complexity and glossing had a significant impact on target GJT scores, logit mixed-effects models were constructed with R. The dependent variable was GJT scores. The null models contained Subject and Item as random effects. To the null models, Time was added and tested in order to examine whether there were significant changes in the GJT scores over the repeated measurements. The results of likelihood ratio tests revealed that the participants' GJT scores increased significantly for both the immediate: $\chi^2(1) = 10.04$, $p < .01$, $C = .80$, and the delayed posttest: $\chi^2(1) = 19.88$, $p < .01$, $C = .82$. Then, to the models containing Time as an existing fixed effect, Complexity and Glossing was added one by one and tested in order to explore whether any of them had a significant interaction with Time. As summarized in Table 16, a significant improvement in the model fit was found when Glossing was added to the delayed GJT data.

Table 16. Summary of likelihood ratio tests for predictors on GJT gain scores for target verbs

		χ^2	*df*	*p*	*C*
Pretest ~ immediate posttest gain	Complexity	.34	2	.85	.81
	Glossing	.11	2	.95	.81
Pretest ~ delayed posttest gain	Complexity	2.40	2	.30	.81
	Glossing	7.13	2	.03*	.81

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

Next, a post-hoc logit mixed effects model was constructed including Glossing and Time as fixed effects, and the results are presented in Table 17. Significant effects for Time were found, implying that learning of the target verbs had occurred in a delayed posttest. The effects of Glossing on GJT gain scores (i.e., Time*Glossing), however, slightly missed significance. In short, participants in this study gained knowledge of the target verbs, as manifested in their delayed posttest scores, although Complexity and Glossing did not significantly influence the extent of development.

Table 17. Summary of a mixed-effects model for Time and Glossing on GJT gain scores for target verbs

		Fixed effects				Random effects	
						by Subject	by Item
Target		Estimate	*SE*	*z*	*p*	*SD*	*SD*
Pretest ~	Intercept	.02	.23	.10	.92	.31	1.16
delayed posttest	Time	.18	.05	3.60	< .01**	.09	.08
gain	Glossing	-.16	.22	-.74	.46	.62	.33
	Time*Glossing	.18	.10	1.85	.06^{+}	.36	.02

*Formula: GJTtarget ~ Time * Glossing + (Time*Glossing|Subject) + (Time*Glossing|Item); C = .82.*

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

Next, another series of likelihood ratio tests were conducted to identify whether the participants' GJT scores for the novel verbs increased significantly over the repeated measurements. The results of likelihood ratio tests, however, revealed that there was no significant change in the GJT scores either in the immediate posttest: $\chi^2(1) = .35$, $p = .55$, $C = .80$, or the delayed posttest: $\chi^2(1) = 2.11$, $p = .15$, $C = .82$. Complexity or Glossing made a significant difference to the null models as regards GJT gain scores for novel verbs. In short, the participants' knowledge in the novel unaccusative verbs did not significantly improve over the pretest, posttest, and the delayed posttest.

3.2. Effects of Complexity and Glossing on the learning of pseudo-words

Table 18 presents descriptive statistics for vocabulary recognition scores by group. The mean scores on the form recognition test were, overall, higher than those on the meaning recognition test. Also, the mean and form recognition scores from the delayed posttest were higher than those from the immediate posttest, whereas the mean meaning recognition scores on the delayed posttest were lower than those on the immediate posttest.

Table 18. Descriptive statistics for vocabulary recognition scores

			Form		Meaning	
Group	Test	*N*	Mean	*SD*	Mean	*SD*
[– C, – G]	Immediate posttest	13	4.15	2.15	2.46	1.90
	Delayed posttest	13	6.08	2.02	1.92	1.38
[– C, + G]	Immediate posttest	13	6.62	1.71	3.62	1.19
	Delayed posttest	13	7.46	1.90	2.77	1.36
[+ C, – G]	Immediate posttest	13	5.15	2.21	2.46	1.71
	Delayed posttest	13	5.54	2.11	2.23	2.01
[+ C, + G]	Immediate posttest	13	5.85	2.55	3.85	1.77
	Delayed posttest	13	5.69	1.65	3.69	1.18

Note. Maximum score for: form recognition = 10, meaning recognition = 10.

To examine whether Complexity and Glossing improved the null models to a significant degree, repeated likelihood ratio tests were conducted using χ^2 statistics. The dependent variables were scores in the immediate posttest and the delayed posttest. The null models included random effects only (i.e., Subject and Item) and each of the fixed effects (i.e., Complexity and Glossing) was added and tested against the null models. As shown in Table 19, Glossing made a significant difference to the null model in an immediate posttest, whereas Complexity improved the null model in a delayed posttest.

Table 19. Summary of likelihood ratio tests for predictors on vocabulary form recognition scores

		χ^2	*df*	*p*	*C*
Immediate posttest	Complexity	.26	1	.61	.77
	Glossing	4.98	1	.03*	.76
Delayed posttest	Complexity	4.98	1	.03*	.79
	Glossing	2.25	1	.13	.79

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

Then, logit mixed-effects models were constructed with Glossing for the immediate posttest scores, and with Complexity for the delayed posttest scores. As Table 20 shows, Glossing was shown to affect form recognition scores in an immediate posttest, whereas Complexity had significant negative effects in a delayed posttest. The *C* indices of concordance were .76 for the model for the immediate posttest and .79 for the delayed posttest, which indicated a good model fit for the data. In short, participants who read glossed texts were better at recognizing word forms in an immediate posttest than those who read unglossed texts. Also, participants in the +

complex conditions scored significantly less in a delayed vocabulary form recognition test than those in the – complex conditions.

Table 20. Summary of mixed-effects models for Glossing on immediate vocabulary form recognition scores

		Fixed effects				Random effects	
						by Subject	by Item
		Estimate	*SE*	*z*	*p*	*SD*	*SD*
Immediate	Intercept	.17	.20	.86	.39	.40	.51
	Glossing	.56	.25	2.25	.03*	.79	.12
Formula: VF ~ Glossing + (Glossing | Subject) + (Glossing | Item); C = .76.							
Delayed	Intercept	.59	.24	2.48	.01*	.45	.62
	Complexity	-.63	.27	-2.30	.02*	.93	.06
Formula: VF ~ Complexity + (Complexity | Subject) + (Complexity | Item); C = 79.							

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

Finally, the effects of task complexity and glossing on vocabulary meaning recognition were explored, beginning with another series of likelihood ratio tests using χ^2 statistics. Again, the null models included only random effects, Subject and Item, and the two fixed effects, Complexity and Glossing, were added to the null models one by one, and it was examined whether this inclusion improved the model fit to a significant extent. As summarized in Table 21, Glossing was shown to make a significant difference to the null models, in both in immediate and delayed posttests.

Table 21. Summary of likelihood ratio tests for predictors on vocabulary meaning recognition scores

		χ^2	*df*	*p*	*C*
Immediate posttest	Complexity	.31	1	.58	.80
	Glossing	6.93	1	.01*	.79
Delayed posttest	Complexity	1.64	1	.20	.83
	Glossing	7.85	1	.01*	.82

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

With Glossing, best-fit models were produced starting with a maximal random effects structure. As displayed in Table 22, Glossing seemed to facilitate recognition of target word meanings in both immediate and delayed posttests. The model fit was relatively good for the data ($C = .79$ for the immediate, $C = .82$ for the delayed). In sum, participants who read glossed texts were better at recognizing word meanings than those who read unglossed texts.

Table 22. Summary of mixed-effects models for Glossing on vocabulary meaning recognition scores

		Fixed effects				Random effects	
						by Subject	by Item
		Estimate	*SE*	*z*	*p*	*SD*	*SD*
Immediate	Intercept	-1.12	.33	-3.42	< .01**	.48	.91
	Glossing	.96	.34	2.81	.01*	.95	.42
Formula: VM ~ Glossing + (Glossing | Subject) + (Glossing | Item); C = .79.							
Delayed	Intercept	-1.50	.41	-3.63	< .01**	.58	1.15
	Glossing	1.34	.46	2.94	.01*	1.17	.76
Formula: VM ~ Glossing + (Glossing | Subject) + (Glossing | Item); C = .82.							

Note. VM = vocabulary meaning recognition scores; Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

Again, likelihood ratio tests were performed to examine if vocabulary recognition scores were influenced by interaction of Complexity and Glossing, and the results revealed that there was no significant interaction effect, Form: $\chi^2(1) = 1.77$, $p = .18$, $R^2 = .02$, Meaning: $\chi^2(2) = .32$, $p = .57$, $R^2 = .03$.

3.3. Interim summary

In sum, neither task complexity nor glossing affected participants' reading comprehension or GJT scores. Effects for task complexity and glossing were found in the vocabulary recognition scores. More specifically, Complexity hindered form recognition of target pseudo-words in a delayed posttest, and glossing facilitated meaning recognition in immediate and delayed posttests. The following section examines the nature of the knowledge acquired through analysis of various measures of implicit and explicit knowledge (Rebuschat, 2013).

4. Source and nature of learned knowledge

In order to explore the source and nature of acquired knowledge, reaction times, binary confidence ratings and subjective source attributions were analysed.

4.1. Reaction times for grammaticality judgment tests

As shown in Table 23, the reaction times appeared to decrease from a pretest to immediate and delayed posttests. It was also observed that, overall, it took longer for the

participants to respond to ungrammatical sentences than to grammatical ones. Paired-sample *t*-tests confirmed that it took significantly longer for participants to respond to ungrammatical sentences than to grammatical sentences in a pretest: $t(51) = -2.41$, $p = .02$, 95% CI [-1426.80, -128.85]. Cohen's *d* was .37, which was evaluated as a small effect size. Significance was slightly missed in an immediate posttest: $t(51) = -2.00$, $p = .05$, 95% CI [-650.22, .96] and a delayed posttest: $t(51) = -1.94$, $p = .06$, 95% CI [-506.02, 9.99]. Cohen's *d*s were .21 and .19, respectively, which indicated small effect sizes.

Table 23. Average reaction time for GJT (milliseconds)

			Grammatical		Ungrammatical	
Group	Test	*N*	Mean	*SD*	Mean	*SD*
[– C, – G]	Pretest	13	6,331	1,613	6,896	1,890
	Immediate posttest	13	4,603	1,092	4,851	1,347
	Delayed posttest	13	3,869	1,222	4,362	1,146
[– C, + G]	Pretest	13	5,804	1,457	6,591	744
	Immediate posttest	13	5,194	1363	5,534	1,179
	Delayed posttest	13	4,073	1,042	4,700	1,209
[+ C, – G]	Pretest	13	6,065	2,170	6,510	2,593
	Immediate posttest	13	5,082	1,424	5,503	2,129
	Delayed posttest	13	4,959	1,479	4,634	1,327
[+ C, + G]	Pretest	13	5,669	1,888	6,985	4,295
	Immediate posttest	13	5,386	1,804	5,676	2,296
	Delayed posttest	13	4,433	1,335	4,632	1,286

Next, in order to examine whether reaction times to the target GJT items changed across the repeated measurements, a series of likelihood ratio tests were conducted on the reaction time data. The null models contained only random effects (Subject and Item), and the increased models additionally included Time as a fixed effect. The results revealed that reaction times decreased significantly from the pretest to the immediate: $\chi^2(1) = 74.17$, $p < .01$, $R^2 = .02$, and the delayed posttest: $\chi^2(1) = 216.3$, $p < .01$, $R^2 = .07$. To the increased models containing Time as an existing fixed effect, Complexity and Glossing was entered one by one and tested whether this inclusion improved the model fit significantly. As Table 24 presents, the results indicated Glossing as a significant factor in the delayed posttest, and Complexity in the immediate posttest. As shown in Table 25, however, the results of post-hoc mixed-

effects modelling showed that neither Complexity nor Glossing had significant influence on the decreasing trend in the reaction time data for the target GJT items.

Table 24. Summary of likelihood ratio tests for predictors on RTs to GJT tests

Target		χ^2	*df*	*p*	R^2
Pretest ~ immediate posttest	Complexity	.25	2	.88	.02
	Glossing	7.80	2	.02*	.03
Pretest ~ delayed posttest	Complexity	14.21	2	< .01**	.08
	Glossing	.76	2	.68	.07
Novel					
Pretest ~ immediate posttest	Complexity	2.12	2	.35	.06
	Glossing	.66	2	.72	.06
Pretest ~ delayed posttest	Complexity	10.48	2	.01*	.11
	Glossing	.53	2	.77	.11

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

Table 25. Summary of mixed-effects models for interaction among Time, Complexity and Glossing on reaction times to GJT items

	Fixed effects			Random effects	
				by Subject	by Item
	Estimate	*SE*	*t*	*SD*	*SD*
Intercept	8.81	.09	102.84°	.46	.24
Time	-.18	.04	-4.69°	.20	–
Glossing	-.18	.15	-1.24	.31	–
Time*Glossing	.13	.08	1.65	.24	–
*Formula: log(RT) ~Time*Glossing + (Time*Glossing | Subject) + (1 | Item); R^2 = .03.*					
Intercept	8.79	.07	121.18°	.29	.21
Time	-.16	.02	-6.84°	.11	–
Complexity	-.10	.13	-.80	.61	–
Time*Complexity	.07	.05	1.43	.20	–
*Formula: log(RT) ~Time*Complexity + (Time*Complexity | Subject) + (1 | Item); R^2 = .07.*					
Intercept	8.90	.07	136.81°	.11	.19
Time	-.18	.01	-13.26°	–	–
Complexity	-.17	.09	-1.93	.41	–
Time*Complexity	.09	.03	3.24°	–	–
*Formula: log(RT) ~Time*Complexity + (Complexity | Subject) + (1 | Item); R^2 = .07.*					

Note. RT = Reaction times; Significance: °$|t| > 2.0$.

The same procedure was conducted on the reaction times data to the novel GJT items. Again, Time emerged as a significant factor in the likelihood ratio tests in the immediate: $\chi^2(1) = 85.51$, $p < .01$, $R^2 = .06$, and the delayed posttest: $\chi^2(1) = 161.71$, $p < .01$, $R^2 = .11$. In other words, reaction times decreased significantly over the repeated measurements. When Complexity and Glossing was added to the models one by one, as shown in Table 24, Complexity improved the model fit significantly for the changes in the reaction times in the delayed posttest. The post-hoc mixed-effects modelling

confirmed that Complexity had a significant influence on the decreasing trend in the reaction times data for the novel GJT items in the delayed posttest.

4.2. Reaction times for vocabulary recognition tests

As shown in Table 26, reaction times in vocabulary form recognition tests showed that they decreased in unglossed conditions, but increased in glossed conditions. In the case of meaning recognition, reaction times decreased in all conditions. As done with reaction time data for the GJT, a series of paired-sample *t*-tests revealed that meaning recognition took significantly longer than form recognition in both immediate, $t(45) = -9.81$, $p < .01$, 95% CI [-2451.79, -1616.58] and delayed posttests, $t(43) = -2.80$, $p = .01$, 95% CI [-1493.86, -241.88]. Cohen's *d*s were .80 and .87, respectively, which were considered small effect sizes.

Table 26. Average reaction time for vocabulary recognition (milliseconds)

		Form			Meaning		
Group	Test	*N*	Mean	*SD*	*N*	Mean	*SD*
[– C, – G]	Immediate posttest	13	1,861	614	13	3,981	1,855
	Delayed posttest	13	1,771	486	13	2,747	960
[– C, + G]	Immediate posttest	13	1,586	388	13	3,630	1,331
	Delayed posttest	13	1,677	479	13	3,006	1,128
[+ C, – G]	Immediate posttest	13	1,921	849	13	4,281	1,540
	Delayed posttest	13	1,807	1,107	13	2,884	1,137
[+ C, + G]	Immediate posttest	13	1,899	610	13	3,560	1,544
	Delayed posttest	13	2,312	3,290	13	2,258	683

Note. Missing values were excluded analysis by analysis.

Again, to examine whether Complexity and Glossing had any effects on reaction times in vocabulary recognition tests, likelihood ratio tests were conducted using χ^2 statistics. The null models included only random effects (i.e., Subject and Item) and fixed effects (i.e., Complexity and Glossing) were entered to the null models one by one and tested against the null models. As Table 27 presents, neither Complexity nor Glossing was found to have a significant influence on the reaction time data for the vocabulary form and meaning recognition tests.

Table 27. Summary of likelihood ratio tests for predictors on RTs to vocabulary recognition tests

Form		χ^2	*df*	*p*	R^2
Immediate ~ delayed Posttests	Complexity	.04	1	.84	.00
	Glossing	.39	1	.52	.00
Meaning					
Immediate ~ delayed Posttests	Complexity	.83	1	.36	.01
	Glossing	3.41	1	.07	.02

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

4.3. Confidence ratings for grammaticality judgment tests

Prior to analysing participants' confidence ratings for the GJT, gain scores in each condition were tested against zero. As presented in Table 28, gain scores were, in general, significantly greater than zero for delayed posttests, except for those for the immediate target items in the [+ complex, + glossing] condition. Cohen's *d*s ranged from 1.55 to 2.71, which were evaluated as large effect sizes.

Table 28. Significance of gain scores for GJT and *d'* values

		Mean	*t*	*d'*	*t*
[– C, – G]	Immediate_target	1.77	1.263		
	Delayed_target	1.15	.96		
	Immediate_novel	.15	.55		
	Delayed_novel	.46	.95		
[– C, + G]	Immediate_target	1.54	1.52		
	Delayed_target	4.08	3.94**	.42	2.76*
	Immediate_novel	-.39	-.50		
	Delayed_novel	.85	.97		
[+ C, – G]	Immediate_target	1.53	1.34		
	Delayed_target	2.92	5.85**	-.01	-.02
	Immediate_novel	1.15	1.60		
	Delayed_novel	1.54	3.99**	.07	.22
[+ C, + G]	Immediate_target	3.69	5.01**	.06	.44
	Delayed_target	5.85	6.91**	.62	2.90*
	Immediate_novel	.31	.47		
	Delayed_novel	.85	1.39		

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

Next, for each GJT datum where significant gains were found, sensitivity index *d'* was calculated using a technique developed by Kunimoto et al. (2001). The rationale for using participants' subjective confidence rating was that, if participants have no awareness, there should be no relationship between their confidence and performance. By contrast, if participants do have awareness, their higher confidence should be

associated more with correct responses than with incorrect ones. However, participants might not be reliable in rating their confidence in a series of responses. Binary confidence ratings allow researchers to overcome this limitation by assigning only two options, *high* versus *low* confidence, and analysing binary ratings using *Signal Detection Theory*, a model for separating bias from sensitivity (Green & Swets, 1966). The analysis begins with categorizing participants' responses according to a combination of their confidence levels and the correctness of their responses. More specifically, as Table 29 shows, *hit* signifies when participants report high confidence and their responses are correct, whereas *false alarm* corresponds to high confidence when the response is incorrect. Also, *correct rejection* indicates participants' low confidence when their responses are incorrect, whereas *miss* corresponds to low confidence for a correct response. As expected, *hit* and *correct rejection* imply participants' awareness, whereas *false alarm* and *miss* suggest unawareness. Based on this categorization, sensitivity index *d*', "the distance between the means of the distributions representing correct responses and incorrect responses" (Kunimoto et al., 2001, p. 303), can be easily calculated using the metric provided by Signal Detection Theory. As what matters in signal detection analysis is the distance between means, each participant's bias or varying sensitivity to reporting either how or low confidence can be accounted for. A *d'* of, or below, zero indicates no awareness, whereas a positive *d'* signifies awareness.

Table 29. Categorization for signal detection analysis

	Confidence	
Accuracy	High	Low
Correct	Hit	Miss
Incorrect	False alarm	Correct rejection

As shown in Table 28, *d'* was significantly higher than zero for delayed target items in the [– complex, + glossing] and the [+ complex, + glossing] conditions. In other cases where significant gains were found, *d'* was not significantly different from

zero. That is, overall, participants appeared to be confident in their grammaticality judgments in the glossed conditions. Cohen's *d*s were 1.08 and 1.14, which indicated large effect sizes.

4.4. Confidence ratings for vocabulary recognition tests

Gain scores in the vocabulary recognition tests were significantly greater than zero for all experimental conditions. As shown in Table 30, *d'* values for form recognition were mostly significantly greater than zero. In contrast, *d'* values for meaning recognition were generally found to be significantly larger than zero only for the delayed posttest in the [– complex, + glossing] condition (see Table 31). In short, the participants were, in general, aware of the correctness of their responses in form recognition tests, but unaware in meaning recognition tests.

Table 30. Significance of gain scores and *d'* values for vocabulary form recognition

Group	Test	Mean	*t*	*d'*	*t*
[– C, – G]	Immediate posttest	4.15	6.95**	1.41	3.18**
	Delayed posttest	6.08	10.85**	1.56	4.49**
[– C, + G]	Immediate posttest	6.62	13.95**	1.57	3.43**
	Delayed posttest	7.46	14.17**	.47	1.90
[+ C, – G]	Immediate posttest	5.44	15.30**	.67	1.41
	Delayed posttest	6.19	9.48**	1.35	2.76*
[+ C, + G]	Immediate posttest	5.85	8.28**	.91	1.94
	Delayed posttest	5.69	12.42**	1.45	3.22**

Note. Significantly above zero: * $p < .05$, ** $p < .01$.

Table 31. Significance of gain scores and *d'* values for vocabulary meaning recognition

Group	Test	Mean	*t*	*d'*	*t*
[– C, – G]	Immediate posttest	2.46	4.68**	-1.24	-2.33
	Delayed posttest	1.92	5.02**	-.35	-.74
[– C, + G]	Immediate posttest	3.62	10.93**	.57	1.53
	Delayed posttest	2.77	7.32**	.94	2.73*
[+ C, – G]	Immediate posttest	2.46	5.18**	-.51	-1.03
	Delayed posttest	2.23	4.01**	.21	.45
[+ C, + G]	Immediate posttest	3.85	7.83**	-.19	-.53
	Delayed posttest	3.69	11.26**	.03	.09

Note. Significantly above zero: * $p < .05$, ** $p < .01$.

4.5. Source attribution for grammaticality judgment tests

An analysis of source attribution was conducted for cases where GJT gain scores were significantly larger than zero, indicating development in the knowledge of unaccusative verbs. Participants reported that their grammaticality judgments were mostly based on intuition and rules (see Table 32). The accuracy rates further revealed that the participants were more likely to respond correctly when their judgments were based on memory and rules. In the [+ complex, + glossing] condition, though, grammaticality judgments based on guess were significantly higher than chance. In sum, participants attributed their grammaticality judgments to both conscious (memory and rules) and unconscious (guess and intuition) knowledge, and their judgments were more likely to be correct when the decisions were made based on memory or rules.

Table 32. Mean proportions and mean accuracy rates across source distribution

			Guess	Intuition	Memory	Rule
[– C, + G]	Delayed_target	Proportion	.16	.32	.19	.48
		Accuracy	.66	.59	.70*	.68*
[+ C, – G]	Delayed_target	Proportion	.12	.30	.23	.35
		Accuracy	.58	.57	.72*	.64
	Delayed_novel	Proportion	.10	.29	.22	.39
		Accuracy	.66	.61	.71	.71*
[+ C, + G]	Immeidate_target	Proportion	.20	.27	.17	.36
		Accuracy	.60	.55	.54	.65*
	Delayed_target	Proportion	.21	.21	.20	.39
		Accuracy	.68*	.66	.66	.81**

Note. Significance from .50: * $p < .05$, ** $p < .01$.

4.6. Interim summary

The analysis of reaction times for the GJT data revealed that it took significantly longer for the participants to judge ungrammatical sentences than grammatical ones, indicating they had some knowledge of English unaccusative verbs. Also, it was found that reaction times were significantly longer for a vocabulary meaning recognition test compared to form recognition, presumably implying that differing levels of cognitive processes were required for word form versus meaning recognition. Neither task

complexity nor glossing, however, had any significant effects on the changes in reaction times for the GJT and vocabulary recognition tests.

In addition, an analysis of binary confidence ratings demonstrated that acquired knowledge of target unaccusative verbs might be conscious in the [– complex, + glossing] and the [+ complex, + glossing] conditions, but unconscious in the [+ complex, – glossing] condition. The source attribution data further showed that accuracy rates were, overall, significantly above chance when grammaticality judgments were based on memory or rules. In other words, it appeared that participants might have been more confident and conscious about what they learned about target unaccusative verbs.

As for vocabulary recognition, binary confidence ratings indicated that participants were in general confident about the correctness of their responses in form recognition tests, which was not the case for meaning recognition tests.

5. WMC as a moderator of L2 reading and L2 learning

This section summarizes the results that address RQ (5) to RQ (8), which concern the role of WMC as a covariate factor moderating the effects of task demands and glossing on reading comprehension and the learning of target constructions. More specifically, the following research questions were addressed:

RQ (5) To what extent does working memory capacity moderate the effects of the cognitive demands of second language reading tasks on reading comprehension?

RQ (6) To what extent does working memory capacity moderate the effects of the cognitive demands of second language reading tasks on development in the knowledge of target language constructions?

RQ (7) To what extent does working memory capacity moderate the effects of glossing on second language reading comprehension?

RQ (8) To what extent does working memory capacity moderate the effects of glossing on development in the knowledge of target language constructions?

To examine whether working memory measures tapped into related constructs, Pearson's correlation coefficients were calculated for working memory capacity measures. As presented in Table 33, the results revealed that digit span scores correlated with all other working memory capacity measures. Nonword span scores also shared significant correlation with backward digit span scores and operation span scores. These significant correlations might indicate some common underlying constructs, including the ability to retain information temporarily in short-term memory (Baddeley, 2003a, 2003b). According to Plonsky and Oswald (2014), overall, the sizes of the correlations appeared small, except for the correlation between digit span scores and nonword span scores (.40 ≤ medium < .60).

Table 33. Correlations among working memory capacity indices

		DS	NWS	BDS	OSPAN
DS	Coefficient	1	.50**	.38*	.31*
	Significance		.00	.00	.03
NWS	Coefficient		1	.28*	.29*
	Significance			.05	.04
BDS	Coefficient			1	-.11
	Significance				.48
OSPAN	Coefficient				1
	Significance				

Note. DS = digit span scores, NWS = nonword span scores, BDS = backward digit span scores, OSPAN = operation span scores; Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

To test whether working memory moderated the effects of Complexity and Glossing on reading comprehension scores, likelihood ratio tests were conducted using χ^2 statistics. Null models included Complexity and Glossing as fixed effects, and Subject and Item as random effects, and each working memory measure was entered into the null models one by one to see if inclusion improved the model fit significantly. As shown in Table 34, nonword span scores and backward digit span scores emerged as significant predictors of reading comprehension scores for Text 1, and digit span scores and nonword span scores increased the null models significantly for Text 2.

Table 34. Summary of likelihood ratio tests for interaction among WMC, Complexity and Glossing on reading comprehension scores

		χ^2	*df*	*p*	R^2
Text 1	DS	7.826	4	.10	.02
	NWS	12.130	4	.02*	.02
	BDS	10.814	4	.03*	.02
	OSPAN	1.896	4	.76	.00
Text 2	DS	11.971	4	.02*	.02
	NWS	11.137	4	.03*	.02
	BDS	8.327	4	.08	.02
	OSPAN	3.425	4	.49	.01

Note. DS = digit span scores, NWS = nonword span scores, BDS = backward digit span scores, OSPAN = operation span scores; Significance level: $^{+}p < .1$, $^{*}p < .05$, $^{**}p < .01$.

Table 35 presents the results from linear mixed-effects models that included working memory capacity indices, one by one, in addition to Complexity and Glossing. Whenever the models failed to converge, random parameters were dropped from the one that accounted for the least variance to the next until convergence was achieved. As can be seen in the table, significant interaction was found among nonword span scores, Complexity and Glossing in Text 1. R^2 of the model was .02, indicating a very small effect size.

Table 35. Summary of mixed-effects models for interaction among WMC, Complexity and Glossing on reading comprehension scores

		Fixed effects			Random effects	
					by Subject	by Item
		Estimate	*SE*	*t*	*SD*	*SD*
Text 1	Intercept	.43	.25	1.76	.10	.21
	COM*NWS	.10	.05	.81	–	–
	GL*NWS	.07	.05	1.32	–	–
	COM*GL*NWS	.22	.11	2.11°	–	–
*Formula: RC ~ Complexity * Glossing * NWS + (1 | Subject) + (1 | Item); R^2 = .02.*						
	Intercept	.42	.20	2.12°	.10	.21
	COM*BDS	-.03	.05	-.58	–	–
	GL*BDS	.08	.05	1.72	–	–
	COM*GL*BDS	-.03	.09	-.32	–	–
*Formula: RC ~ Complexity * Glossing * BDS + (1 | Subject) + (1 | Item); R^2 = .02.*						
Text 2	Intercept	.61	.14	4.39°	.07	.11
	COM*DS	-.05	.03	-1.43	–	–
	GL*SD	.05	.03	-1.51	–	–
	COM*GL*SD	.05	.07	.71	–	–
*Formula: RC ~ Complexity * Glossing * DS + (1 | Subject) + (1 | Item); R^2 = .02.*						
	Intercept	.47	.17	2.80°	.06	-2.07
	COM*NWS	.07	.04	1.98	–	–
	GL*NWS	-.02	.04	-.61	–	–
	COM*GL*NWS	.06	.07	.75	–	–
*Formula: RC ~ Complexity * Glossing * NWS + (1 | Subject) + (1 | Item); R^2 = .02.*						

Note. COM = Complexity, GL = Glossing, DS = digit span scores, BDS = backward digit span scores, NWS = nonword span scores; Significance: °| *t* | > 2.0.

Next, post hoc mixed-effects modelling was conducted to examine the differential contribution made by working memory to reading comprehension scores across different experimental conditions. As Table 36 demonstrates, nonword span scores were found to play a significant role in the [+ complex, + glossing] condition for Text 1. R^2 was .08, which was a small effect size. In sum, when assigned in the [+ complex, + glossing] condition, participants with higher nonword span scores were better at answering reading comprehension items for Text 1 than those with lower nonword span scores.

Table 36. Summary of post-hoc mixed-effects models for interaction among NWS, Complexity and Glossing on reading comprehension scores

		Fixed effects			Random effects	
					by Subject	by Item
		Estimate	*SE*	*t*	*SD*	*SD*
Text 1	Intercept	.83	.43	1.96	.30	.45
- COM, - GL	NWS	-.00	.05	-.00	.05	.03
	Formula: RC ~ NWS + (NWS |Subject) + (NWS |Item); R^2 = .00.					
- COM, + GL	Intercept	1.45	.43	3.35°	.23	.65
	NWS	-.08	.05	-1.58	.00	.08
	Formula: RC ~ NWS + (1|Subject) + (NWS |Item); R^2 = .00.					
+ COM, - GL	Intercept	.59	.70	.84	.16	.52
	NWS	.02	.73	.25	.16	.46
	Formula: RC ~ NWS + (1|Subject) + (NWS |Item); R^2 = .00.					
+ COM, + GL	Intercept	-.94	.53	-1.78	.51	.12
	NWS	.19	.06	3.29°	.04	.02
	Formula: RC ~ NWS + (1|Subject) + (1|Item); R^2 = .08.					

Note. COM = Complexity, GL = Glossing, NWS = nonword repetition scores;
Significance: °| t | > 2.0.

The role of working memory capacity as a moderator of the effects of Complexity and Glossing was also investigated in relation to GJT scores. Likelihood ratio tests were conducted to compare reduced models that included Time, Complexity and Glossing as fixed effects with increased models that contained each of the working memory capacity indices as an additional fixed effect. As summarized in Table 37, only operation span scores emerged as a significant predictor of delayed gain scores for the target verbs. Accordingly, mixed-effects models were constructed with operation span scores.

Table 37. Summary of likelihood ratio tests for interaction among WMC, Complexity and Glossing on GJT gain scores

Target		χ^2	*df*	*p*	*C*
Pretest ~ immediate posttest	DS	6.36	8	.61	.81
	NWS	9.24	8	.32	.81
	BDS	4.01	8	.86	.81
	OSPAN	4.55	8	.81	.81
Pretest ~ delayed posttest	DS	10.15	8	.26	.81
	NWS	3.54	8	.90	.80
	BDS	6.30	8	.64	.80
	OSPAN	5.20	8	.01**	.80
Novel					
Pretest ~ immediate posttest	DS	8.56	8	.38	.81
	NWS	5.55	8	.70	.81
	BDS	4.20	8	.84	.81
	OSPAN	10.88	8	.21	.81
Pretest ~ delayed posttest	DS	6.08	8	.64	.83
	NWS	7.42	8	.49	.83
	BDS	5.66	8	.69	.83
	OSPAN	5.57	8	.70	.82

Note. DS = digit span scores, NWS = nonword span scores, BDS = backward digit span scores, OSPAN = operation span scores; Significance level: $^{+}p < .1$, $^{*}p < .05$, $^{**}p < .01$.

A post hoc mixed-effects model was produced to examine the nature of the interaction among Time (i.e., GJT gain scores), Complexity, Glossing and operation span scores. As displayed in Table 38, however, operation span scores did not share significant interaction with either Complexity or Glossing in the model. In other words, working memory capacity did not moderate the effects of Complexity and Glossing on GJT gain scores.

Table 38. Summary of mixed-effects models for interaction among OSPAN, Complexity, and Glossing on target GJT gain scores

		Fixed effects				Random effects	
						by Subject	by Item
Target		Estimate	*SE*	*z*	*p*	*SD*	*SD*
Pretest ~	Intercept	-1.37	1.93	-.71	.48	.61	1.22
delayed	Time*COM*OSPAN	.00	.02	.06	.96	–	–
posttest	Time*GL*OSPAN	-.00	.02	-.08	.93	–	–
	Time*COM*GL*OSPAN	.02	.04	.39	.70	–	–

*Formula: GJTtarget ~ Time * Complexity * Glossing * DS + (1 | Subject) + (1 | Item); C = .80.*

Note. COM = Complexity, GL = Glossing, OSPAN = operation span scores; Significance: $^{+}p < .1$, $^{*}p < .05$, $^{**}p < .01$.

Last but not least, likelihood ratio tests were conducted to identify working memory capacity measures that had moderating effects on vocabulary recognition scores. Null models consisted of Subject and Item as random effects, and Complexity

and Glossing as fixed effects. Increased models contained working memory capacity measures, added one by one, and these were tested against the null models to see if the inclusion of working memory capacity measures improved the model fit significantly. As summarized in Table 39, however, none of the measures emerged as significant predictors. In sum, working memory capacity did not emerge as a significant moderator of the effects of Complexity or Glossing on recognition of the forms and meanings of target pseudo-words.

Table 39. Summary of likelihood ratio tests for interaction among WMC, Complexity and Glossing on vocabulary recognition scores

Form		χ^2	*df*	*p*	*C*
Immediate posttest	DS	1.17	8	.88	.76
	NWS	1.64	8	.80	.76
	BDS	.96	8	.92	.76
	OSPAN	1.95	8	.74	.75
Delayed posttest	DS	2.91	8	.57	.79
	NWS	.76	8	.94	.79
	BDS	4.20	8	.38	.79
	OSPAN	9.06	8	$.06^{+}$	.79
Meaning					
Immediate posttest	DS	3.60	8	.46	.75
	NWS	2.07	8	.77	.75
	BDS	1.00	8	.91	.76
	OSPAN	7.97	8	.09	.76
Delayed posttest	DS	3.31	8	.51	.78
	NWS	3.30	8	.51	.77
	BDS	1.27	8	.87	.87
	OSPAN	6.41	8	.17	.77

Note. DS = digit span scores, NWS = nonword span scores, BDS = backward digit span scores, OSPAN = operation span scores; Significance level: $^{+}p < .1$, $^{*}p < .05$, $^{**}p < .01$.

V. Interim Discussion

This study has investigated whether task complexity and glossing affected Korean undergraduate students' English reading comprehension and their development in the knowledge of target constructions contained in the texts. The target constructions were English unaccusative verbs and ten pseudo-words. It has also explored whether working memory capacity moderated the effects of task complexity and glossing on development in the knowledge of the target constructions. Task complexity was manipulated by

disarranging paragraphs of each text, based on the understanding that a coherent and clear text structure considerably facilitates reading comprehension (Meyer, 1975, 1985; Meyer & Freedle, 1984; Meyer & Ray, 2011). Glossing was done by providing Korean definitions of the target constructions in the margins of texts. The results of this study (for a summary of significant findings, see Table 40) are discussed in this section.

Table 40. Summary of significant results of Study 1

	Dependent variables	Statistical results
Fixed effects		
Complexity	Delayed word form recognition scores	z= -2.30, p = .02*, C = .79
Glossing	Immediate word form recognition scores	z= 2.25, p = .03*, C = .76
	Immediate word meaning recognition scores	z= 2.81, p = .01*, C = .79
	Delayed word meaning recognition scores	z= 2.94, p = .01*, C = .82
Moderator		
NWS	Reading comprehension for Text 1 in [+ C, + G]	t = 3.29°, R^2 = .08

Note. NWS = nonword span scores; Significance: °| t | > 2.0; ^{+}p < .1, *p < .05, **p < .01.

1. Effects of task complexity and glossing on L2 reading comprehension

In this study, reading comprehension scores were affected by neither task complexity nor glossing of the texts. It should however be noted that, as shown in the relatively high mean scores, overall, participants performed well in the reading comprehension tests, and thus a ceiling effect might have masked between-group differences. More difficult reading comprehension tests may result in greater variances among participants, increasing the reliability of reading comprehension tests. In addition, more demanding reading comprehension items may better encourage learners to look up glosses in order to achieve more accurate comprehension of a given text. What should be also noted is that participants were allowed to stay on the task as long as they felt necessary, which might have contributed to the non-significant effects of task manipulation on reading comprehension scores. It seems equally possible that participants' reading processes were, in fact, affected by task complexity or glossing, but these effects did not surface in the reading comprehension scores. To explore this

issue, verbal reports revealing the internal processes of participants while performing tasks, or eye-movement data, would be highly informative.

2. Effects of task complexity on development in the knowledge of target constructions

Task complexity was not found to make a significant contribution to knowledge of English unaccusative verbs. It is possible that the paragraph-ordering task in the + complex conditions did not encourage participants to process target unaccusative verbs to a significantly greater extent. More specifically, the ordering task might have led participants to depend on the initial or final part of each paragraph selectively, rather than paying attention to each paragraph thoroughly. Also, arranging paragraphs might have promoted conceptual reasoning rather than text-based processing, thereby not affecting learning of the target verbs. In other words, when rearranging the paragraphs, participants might have focused more on the main idea of each paragraph and tried to figure out a logical order among the ideas. In addition, the task in the + complex conditions in this study might not have been much more demanding than the – complex task. Indeed, task manipulation was shown to be successful only via subjective time estimations, not by self-reports on the perceived mental effort (for the validity of self-reports in assessing task complexity, see Révész, Michel, & Gilabert, 2016).

When it comes to the incidental learning of target pseudo-words, task complexity was shown to have significant negative effects on form recognition in a delayed posttest. That is, participants assigned to the + complex conditions were less successful in recognizing target word forms than those in the – complex conditions. It seems possible that the increased level of task complexity could have directed participants' attention to the paragraph-ordering task, and thus away from attending to pseudo-word forms. When participants were allowed to read the text in a coherent order under the – complex conditions, they might have employed extra mental resources to be shared out

for processing target word forms. This is open to empirical investigation, preferably using online methodologies such as concurrent or retrospective verbal reports or eye-movement data.

3. Effects of glossing on development in the knowledge of target constructions

This study has also demonstrated that glossing did not affect knowledge of target unaccusative verbs. The effects of glossing in the delayed posttest, however, approached significance ($z = 1.85$, $p = .06$, $C = .82$), which calls for a follow-up study with a more rigorous research design. Another ground for a follow-up study is the unique characteristics of English unaccusative verbs. In Guidi's (2009) and Martinez-Fernández's (2010) studies, where glossing failed to affect learning of L2 grammatical features, the target constructions were the Spanish subjunctive and present perfect, which involve complex conjugations. English unaccusativity, however, can be seen as a construction that entails lexical patterns and item-based learning (Lee et al., 2008; Sorace, 2000; Zyzik, 2009), and hence could be more susceptible to glossing. In Nagata's (1999) study where significant effects of glossing were found, the target grammatical structures (i.e., a conjunction, *hodo*, and two postpositional particles, *kara* and *to*) also involved lexical patterns. Given that the effects of glossing may interact with the type of target linguistic item (Guidi, 2009), more research into how glosses affect the learning of diverse types of L2 grammatical features could generate insights into the combined effects of glossing and the nature of TL constructions.

In this study, glossing facilitated receptive knowledge of target pseudo-words, as evidenced by the immediate form recognition scores and immediate and delayed meaning recognition scores. The long-term effects of glossing on word meaning recognition, along with longer reaction times and relatively lower confidence ratings compared with form recognition, seem to reflect the complex cognitive processes involved in meaning recognition. That is, word meaning recognition requires not only

identifying visual patterns of perceived letters, which suffices in the case of form recognition, but also engaging in phonemic recoding from registered sequences of letters and finding their semantic qualities from a lexical inventory. Accordingly, the meaning provided in glosses might have been more deeply processed than word form, resulting in more robust retention (Craik & Lockhart, 1972; Hulstijn & Laufer, 2001).

4. WMC as a mediator of the effects of task complexity and glossing

In this study, working memory capacity had moderating effects only on reading comprehension scores. More specifically, when reading glossed texts under the + complex condition, participants appeared to benefit from higher nonword span when doing reading comprehension items. The results may indicate that, phonological short-term memory plays an important role in retaining glosses or multiple propositional units. This finding is noteworthy, given that most previous studies have focused on the role of complex working memory in L2 reading comprehension, predominantly using a reading span task, while the role of phonological short-term memory has been largely ignored (e.g., Alptekin & Erçetin, 2009, 2011; Alptekin et al., 2014; Leeser, 2007; Walter, 2004). With respect to the lack of moderating effects of working memory capacity on L2 learning scores, the role of working memory capacity could have operated at the level of noticing, although it did not become apparent in the test scores. In this regard, more research should explore whether working memory moderates the effects of task complexity and glossing on the noticing of L2 constructions.

VI. Insights for Study 2

The results from Study 1 offered valuable insights for Study 2. First, it was speculated that a ceiling effect could have masked the effects of task complexity or glossing on reading comprehension scores. Indeed, mean scores were high while variances were small, suggesting an inherent limitation in detecting significant effects

of task complexity and glossing on reading comprehension scores. Therefore, it was assumed that the difficulty of reading comprehension tests might need to be increased in Study 2, so that variances among the participants' scores could be inflated. It was also suggested that, to better detect the effects of task complexity, task manipulation should be conducted on a more localized level. In this study, the tasks were manipulated on a global level (rearranging paragraphs into a coherent order) and thus this led the participants to rely on top-down conceptual processing, which, in turn, might have failed to affect the linguistic processing of target constructions. It was speculated that local-level task manipulation might encourage learners to read a given text more thoroughly so that target features could be processed more deeply in the + complex task conditions. It was also considered that a time limit might play a role in magnifying the effects of the cognitive complexity of each task by placing additional cognitive demands on the participants.

CHAPTER 4

STUDY 2

Study 2 replicated Study 1, but on a larger scale, while employing the same research design. The limitations discussed in Study 1 were considered when modifying the research instruments so that the reliability and validity of the results could be enhanced. The major modification made in Study 2 was the way task demands were manipulated, and only minor or no changes were made to other research instruments. This selective and focused modification to the research design was expected to allow the researcher to compare and contrast the results from Study 1 and Study 2 in a more systematic way, and thereby identify the source of different findings more accurately. For the reader's convenience, the research questions are repeated here.

RQ (1) To what extent do the cognitive demands of second language reading tasks affect reading comprehension?

RQ (2) To what extent do the cognitive demands of second language reading tasks affect development in the knowledge of target language constructions?

RQ (3) To what extent does glossing affect second language reading comprehension?

RQ (4) To what extent does glossing affect development in the knowledge of target language constructions?

RQ (5) To what extent does working memory capacity moderate the effects of the cognitive demands of second language reading tasks on reading comprehension?

RQ (6) To what extent does working memory capacity moderate the effects of the cognitive demands of second language reading tasks on development in the knowledge of target language constructions?

RQ (7) To what extent does working memory capacity moderate the effects of glossing on second language reading comprehension?

RQ (8) To what extent does working memory capacity moderate the effects of glossing on development in the knowledge of target language constructions?

This chapter first describes the methodology of the present study, including the research design, the participants, and the research instruments, while highlighting the differences from Study 1. The next section presents the results of mixed-effects modelling with the obtained data, followed by a summary of findings and interim discussions that lead to Study 3.

I. Research Design and Methodology

1. Design

As shown in Figure 14, a pretest, posttest and delayed posttest design was employed with 88 participants who were randomly assigned to one of four experimental conditions, i.e., [+ complex task, + glossing], [– complex task, + glossing], [+complex task, – glossing] and [– complex task, – glossing]. In each treatment session, participants read a TOEFL passage while answering multiple-choice reading comprehension items. Scores from the reading comprehension items were used as an index for reading comprehension, and development in the knowledge of the target constructions was measured with a grammaticality judgment test and a vocabulary recognition test. In the last session, the participants' working memory capacity was measured with the same memory span tasks used in Study 1. Also, various questionnaires were administered in order to obtain information regarding the participants' previous English learning experiences and their perceptions about the reading tasks, such as topic familiarity and task difficulty.

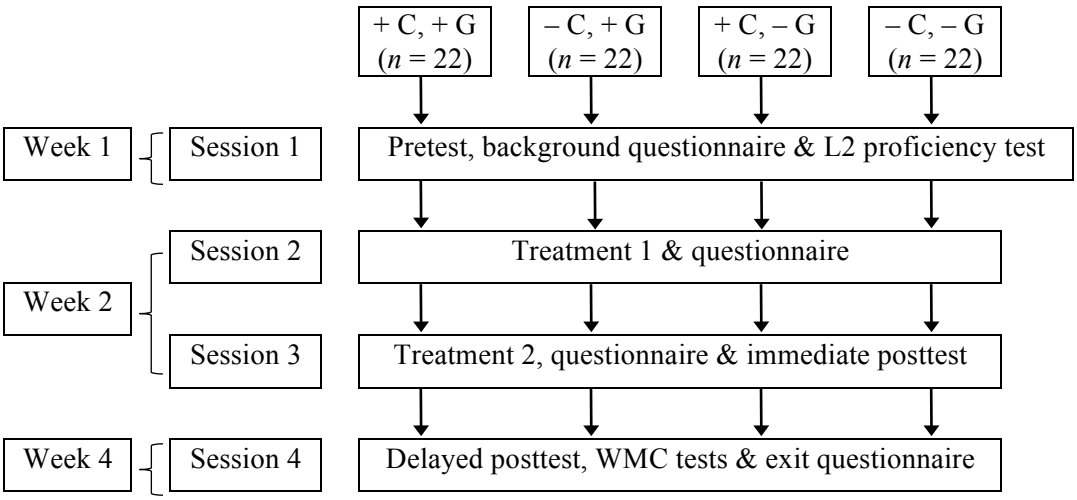

Figure 14. Experimental design and procedure for Study 2

2. Participants

The participants in Study 2 comprised 88 (53 male and 35 female) undergraduate students enrolled at a university in Korea. Their L1 was Korean and they had no experience of living in an English-speaking country before the age of 20. The average age of the participants was 23.69 years (*SD* = 3.67), and the average onset age of English learning was 9.45 years (*SD* = 2.52). Nine students had experience of living in English-speaking countries, including the US, Canada, Australia and the Philippines (Mean = 7.33 months, *SD* = 7.89 months). To ensure the homogeneity of participants' English ability, their English proficiency level was measured with an adapted version of the *Reading and Use of English* section of a practice *Cambridge Proficiency: English* (CPE) test developed and provided by *University of Cambridge ESOL examinations*. The CPE test was modified as each treatment session had to be completed within an hour due to limited time allowed for data collection. Thus, items that lowered test reliability were removed, resulting in 30 items in total. Based on the CPE scores, stratified random sampling was applied in order to ensure equivalence among the groups in terms of English proficiency.

3. Materials

While Study 2 replicated Study 1, a few modifications were made to overcome some limitations observed in Study 1 and, in so doing, enhance the internal validity of Study 2.

3.1. Target constructions

The two texts used in Study 1 were also used as treatment texts in Study 2. While the same English unaccusative verbs were included as target constructions, *decompose* and *fossilize* were removed because native speakers exhibited different opinions regarding the correct usage of those verbs in either active or passive voices. As illustrated in Table 41, six verbs were non-alternating unaccusative verbs, whereas others were alternating ones. Each of the unaccusative verbs appeared once in the texts.

Table 41. Target English unaccusative verbs for Study 2

	Text 1				Text2		
	Unaccusative verbs	Alterna -ting	Frequency (per 450 million)		Unaccusative verbs	Alternating	Frequency (per 450 million)
1	subside	NA	568	1	date to	A	743
2	ascend	A	759	2	originate	A	1,022
3	accumulate	A	1,814	3	consist of	NA	2,140
4	cease	A	2,554	4	persist	NA	2,684
5	diminish	A	2,701	5	evolve	A	3,184
6	drift	NA	4,477	6	disappear	NA	7,581
7	collect	A	10,525	7	emerge	NA	9,116
8	settle	A	10,873				

There were a few changes to the target pseudo-words as well. As the extent to which correct meaning could be inferred from the context was considered different across the words in Study 1, three Korean speakers who were experts in Applied Linguistics were invited to assess the level of difficulty/ possibility of inferring correct meanings of the words. Three words (i.e., *klaners* for parks, *stovons* for beaches, and *tralion* for predator) were considered relatively easy or problematic, in that multiple meanings could be allowed, and hence replaced with other nouns (*klenear* for surface,

tralion for seawater, and *stovons* for conditions) that were considered comparable in terms of inferability (see Table 42).

Table 42. Target pseudo-words for Study 2

	Text 1			Text2	
	Pseudo-word	Original word		Pseudo-word	Original word
1	stragon	bottom	1	cabrons	changes
2	golands	spouts	2	fration	absence
3	phosens	discoveries	3	zenters	clues
4	klenear	surface	4	morbits	descendants
5	tralion	seawater	5	stovons	conditions

3.2. Task complexity manipulation

In Study 1, task complexity had limited effects on the participants' perceived level of task difficulty, probably because the task manipulation was conducted on the paragraph/ discourse level. In order to amend this, the tasks were manipulated on a local level in Study 2 (see Figure 15 and 16). More specifically, under the – complex condition, each text segment was split into two subparts (A and B), and participants were asked to determine which came first of the two. Under the + complex condition, each segment was split into three or four parts (A, B, C and D), and participants were asked to re-arrrange them coherently. On top of that, unlike in Study 1, a time limit of 25 minutes was set for task completion, which was expected to augment differences in the cognitive demands between the + and – complex task conditions. From the 14 reading comprehension items used in Study 1, low-reliability items were either deleted or modified, resulting in 9 items in total for each text. The total score for reading comprehension test was 10 (1 point for each of 8 items and 2 points for an item requiring summary completion) for each text. Again, glossing was conducted by providing Korean definitions of target features in the margins of the texts.

Directions: Read the provided passage and answer the comprehension questions. You have 25 minutes.

THE CAMBRIAN EXPLOSION

[A] Animals originated[1] relatively late in the history of Earth – in only the last 10 percent of Earth's history. During a geologically brief 100-million-year period, all modern animal groups (along with other animals that are now extinct) came to exist. This rapid origin and diversification of animals is often referred to as "the Cambrian explosion."

[B] The geologic timescale consists of[2] significant geologic and biological events, including the origin of Earth about 4.6 billion years ago, the origin of life about 3.5 billion years ago, the origin of eukaryotic life-forms (living things that have cells with true nuclei) about 1.5 billion years ago, and the origin of animals about 0.6 billion years ago. The last event marks the beginning of the Cambrian period.

[1] 발생하다, 유래하다

[2] 구성되다

1. First, put [A] and [B] in a correct order.

   _____ → _____ (B-A)

2. The word "significant" in the passage is closest in meaning to
   (a) numerous
   (b) consequential
   (c) unexplained
   (d) unexpected

Figure 15. Sample task layout of – complex condition for Study 2

Directions: Read the provided passage and answer the comprehension questions. You have 25 minutes.

THE CAMBRIAN EXPLOSION

[A] The last event marks the beginning of the Cambrian period. **[B]** The geologic timescale consists of[1] significant geologic and biological events, including the origin of Earth about 4.6 billion years ago, the origin of life about 3.5 billion years ago, the origin of eukaryotic life-forms (living things that have cells with true nuclei) about 1.5 billion years ago, and the origin of animals about 0.6 billion years ago. **[C]** This rapid origin and diversification of animals is often referred to as "the Cambrian explosion." **[D]** Animals originated[2] relatively late in the history of Earth – in only the last 10 percent of Earth's history. During a geologically brief 100-million-year period, all modern animal groups (along with other animals that are now extinct) came to exist.

[1] 구성되다

[2] 발생하다, 유래하다

1. First, rearrange [A] to [D] in a correct order.

   _____ → _____ → _____ → _____ (B-A-C-D)

2. The word "significant" in the passage is closest in meaning to
   (a) numerous
   (b) consequential
   (c) unexplained
   (d) unexpected

Figure 16. Sample task layout of + complex condition for Study 2

3.3. Assessment tasks

The GJTs used for this study included 80 sentences in total (see Appendix B-2). For each of the 15 target unaccusative verbs, one grammatical and one ungrammatical passive sentence were constructed, resulting in 30 sentences. In addition, 16 sentences were constructed with the 8 additional unaccusative verbs to see if learning transferred to other unaccusative verbs that were not contained in the treatment texts. Lastly, 34 distractors, 17 grammatical and 17 ungrammatical, were included. The distractors tapped into the same grammatical features as in the distractor sentences used in Study 1 (No & Chung, 2006). While some sentences were taken from Study 1, most sentences were newly constructed in Study 2, in order to better control variables such as word frequency, semantic plausibility, the number of syllables and the position of unaccusative verbs in each pair of sentences. Three native speakers were invited to review the grammaticality and semantic plausibility of the sentences. As done in Study 1, binary confidence ratings and source attributions were collected for each response. Across a pretest and immediate and delayed posttests, the same 80 sentences were randomly presented to the participants, and a GJT took approximately 10 to 12 minutes to complete.

As there were some changes in the selection of target lexical items to be replaced with pseudo-words, the vocabulary form and meaning recognition tests were revised accordingly, while keeping the overall structure of the tests used in Study 1 (see Appendices D-2 and E-2). As in Study 1, the vocabulary form recognition test contained 20 items, 10 for the target words and 10 for the distractors. Participants were instructed to check either *yes* or *no*, depending on whether they remembered seeing the word in the texts. The vocabulary meaning recognition test was in a multiple-choice format with four options and containing 10 items for the target words and 10 for the distractors. For

both form and meaning recognition tests, binary confidence ratings were collected for each response. The vocabulary recognition test took 6 to 7 minutes.

Last but not least, the same forward digit span test (DS, Cronbach's alpha = .73) and nonword repetition test (NWR, Cronbach's alpha = .79) as used in Study 1 were used to measure participants' phonological short-term memory. Also, the same backward digit span test (BDS, Cronbach's alpha = .84) and automated operation span test (OSPAN, Cronbach's alpha = .78) were employed to measure participants' complex working memory.

3.4. Questionnaires

The questionnaires were largely the same as in Study 1 (see Appendix E). More specifically, participants answered a background questionnaire, post-reading questionnaires and an exit questionnaire. The background questionnaire was to collect participants' demographic information and English learning experience. The post-reading questionnaire asked participants to assign a perceived level of task difficulty and topic familiarity to the reading texts. Unlike in Study 1, participants were not asked to estimate how long they took to complete the tasks retrospectively, as there was a time limit for task completion. Finally, an exit questionnaire asked participants to give retrospective comments on their experience of task performance. All questionnaires were administered in Korean.

II. Procedure

As displayed in Figure 14, data were collected over four weeks in a computer laboratory at a university in Korea. As in Study 1, the participants read an information sheet explaining their rights, the risks and the benefits of participation and signed a consent form at the beginning of the first session (see Appendix A-2). All participants completed a pretest, a background questionnaire and an L2 proficiency test in the first

session. In sessions 2 and 3, the participants took part in treatment sessions, each followed by completion of a post-reading questionnaire. In session 3, they also completed an immediate posttest. In the fourth week, participants completed a delayed posttest and working memory tests. Each session took approximately 45 minutes to an hour.

III. Analysis

As in Study 1, SPSS 22.0 was used to examine the reliability of the tests as well as to compute descriptive and correlational statistics for the data. More specifically, the reliability of the different tests was determined using Cronbach's alpha, and interrelationships between the various test scores were computed using Pearson's coefficient. Mixed-effects models were constructed to compare various mean values among the groups, explore the effects of task complexity and glossing on reading comprehension scores and obtain scores for the learning assessment tasks. In order to do this, the statistical program R, version 3.3.0 was used (R Development Core Team, 2016). For interval-scale data (reading comprehension scores and reaction time data), linear mixed effects models were constructed using *lmer* function, while categorical-scale data (GJT scores and vocabulary recognition scores) were analysed using logit mixed-effects models with *glmer* function provided in the package *lme4* (Bates et al., 2012).

Effect sizes for linear models were evaluated with R^2 using *r.squaredGLMM* function provided by the package *MuMln* (Barton, 2015), whereas those for logit models were assessed with *C* index of the concordance using *somer 2* function in the *Hmisc* package (Harrell & Dupont, 2015). Following Plonsky and Oswald (2014), R^2 values of .06, .16 and .36 were interpreted as small, medium and large, respectively. A *C*-index of .70 was considered a moderate fit, .80 good and .90 and above excellent for

the data (Baayen, 2008). For *t*-tests, Cohen's *d* was calculated to examine effect sizes. The benchmarks were .40 for small, .70 for medium and 1.00 for large effect sizes for independent-sample *t*-tests, whereas .60 for small, 1.00 for medium and 1.50 for large effect sizes for paired-sample *t*-tests (Plonsky & Oswald, 2014). Collinearity statistics for the independent variables were calculated using *collin.fnc* function in the *languageR* package (Belsley et al., 1980). Condition numbers between 0 and 6 were regarded as no collinearity, around 15 as medium, and 30 or above as potentially harmful collinearity (Baayen, 2008).

IV. Results

1. Preliminary analysis

As in Study 1, some preliminary analyses were conducted to assess the reliability and validity of the instruments.

1.1. Test reliability

Table 43 presents descriptive statistics for the proficiency test, the reading comprehension tests, the grammaticality judgment tests and the vocabulary recognition tests. As shown in the table, the mean score was considered very small for the vocabulary meaning recognition test (mean = 3.33 out of 20).

Table 43. Descriptive statistics for test scores

	N	Mean	*SD*	*Cronbach's alpha*
CPE test	88	8.72	3.63	.68
Reading comprehension (Text 1)	88	5.94	2.04	.57
Reading comprehension (Text 2)	88	6.11	1.75	.52
Grammaticality judgment (Target items)	88	49.34	12.33	.83
Grammaticality judgment (Novel items)	88	31.41	7.01	.76
Vocabulary recognition (Form)	88	10.56	3.95	.66
Vocabulary recognition (Meaning)	88	3.33	2.35	.48

Note. Maximum score for: CPE test = 30, reading comprehension = 10, grammaticality judgment (target) = 90, grammaticality judgment (novel) = 48, vocabulary form recognition = 20, vocabulary meaning recognition = 20.

Reliability coefficients for the test scores were computed using Cronbach's alpha. As summarized in the table, the values of Cronbach's alpha were, overall, lower for the reading comprehension tests and vocabulary meaning recognition tests. That said, the results for the reading comprehension scores and vocabulary meaning recognition scores should be interpreted with some caution.

1.2. English proficiency

CPE items that were shown to lower reliability in Study 1 were identified and removed, resulting in 30 items in total. As presented in Table 44, it appeared that the CPE test was quite difficult overall for the participants. To check the equivalence of English proficiency among the groups, a mixed-effects model was constructed for the CPE scores with Group as a fixed effect and Subject and Item as random effects. When compared with a null model that contained only random effects, the results showed that inclusion of Group as a fixed effect did not make a significant difference to the null model, $\chi^2(1) = 29$, $p = .59$, $R^2 < .01$. In other words, the groups did not significantly differ from each other in terms of their English proficiency.

Table 44. Descriptive statistics for CPE scores

	Proficiency test			
Group	*N*	Mean	*SD*	*SE*
[– C, – G]	22	8.36	3.22	.69
[– C, + G]	22	8.55	4.36	.93
[+ C, – G]	22	8.96	3.43	.73
[+ C, + G]	22	9.00	3.63	.77
Total	88	8.72	3.63	.39

Note. Maximum score = 30.

1.3. GJT scores on the pretest

Next, another set of likelihood ratio tests were conducted on the pretest GJT scores, comparing null models only with random effects, and increased models containing Group as a fixed effect to ensure that the four groups started out at a developmentally parallel stage (for descriptive statistics, see Table 49). The results

indicated that Group did not improve the null models to a significant degree, Target items: $\chi^2(1) = .30$, $p = .59$, $R^2 < .01$; Novel items: $\chi^2(1) = 1.87$, $p = .17$, $R^2 < .01$. That is, the results showed there was no significant difference among the groups in their ability to judge the grammaticality of English unaccusative sentences as reflected in their pretest scores.

1.4. Effects of topic familiarity

To confirm the absence of topic knowledge influence, the effects of the participants' familiarity with each topic were measured using post-reading questionnaire items (i.e., Item 6: *I thought the topic of the reading was familiar*, Item 13: *I had some background knowledge of the reading topic*). Descriptive statistics for the responses to the two items are provided in Table 45. The responses to the two items correlated significantly, Text 1: $r(88) = .87$, $p < .01$, Text 2: $r(88) = .72$, $p < .01$. To examine the effects of topic familiarity on reading comprehension scores, likelihood ratio tests were conducted comparing a null model only with random effects and an increased model including topic familiarity as fixed effects, which were summed values of the responses to the two questionnaire items. The results showed that topic familiarity did not make significant difference to the null models, Text 1: $\chi^2(1) = .09$, $p = .77$, $R^2 < .01$; Text 2: $\chi^2(1) = .79$, $p = .38$, $R^2 < .01$. That is to say, the participants' topic familiarity with the texts did not affect their scores on reading comprehension items.

Table 45. Descriptive statistics for topic familiarity by item

		Topic familiarity					
		Text 1			Text 2		
Item	*N*	Mean	*SD*	*SE*	Mean	*SD*	*SE*
# 6	88	3.88	1.69	.18	3.50	1.74	.19
#13	88	3.86	1.73	.18	3.27	1.71	.18
Total	88	7.74	3.31	.35	6.77	3.20	.34

Note. Maximum value for each item = 7.

1.4. Validation of task complexity manipulation

To infer the effects of task complexity on the amount of mental effort imposed on the participants, three questionnaire items were included in post-task questionnaires (Item 1: *I thought this task was difficult*, Item 7: *I invested a large amount of mental effort to complete this task*, Item 14: *I thought this task was demanding*). Descriptive statistics for the responses to the three items are presented in Table 46. Cronbach's alpha for the items was .70 for Text 1 and .76 for Text 2, indicating that the items probably measured overlapping constructs, i.e., the level of mental effort.

Table 46. Descriptive statistics for perceived task difficulty by item

			Reported mental effort					
			Text 1			Text 2		
Item	Condition	*N*	Mean	*SD*	*SE*	Mean	*SD*	*SE*
# 1	– Complex	44	5.09	1.10	.16	4.50	1.39	.21
	+ Complex	44	5.57	.95	.14	5.27	1.00	.15
# 7	– Complex	44	5.43	1.00	.15	5.41	1.00	.15
	+ Complex	44	5.39	1.15	.17	5.55	1.07	.16
# 14	– Complex	44	4.48	1.17	.18	4.23	1.48	.22
	+ Complex	44	5.11	1.28	.19	4.84	1.22	.18
Total	– Complex	44	15.00	3.50	.53	14.14	4.11	.62
	+ Complex	44	16.07	3.03	.46	15.66	2.96	.45

Note. Maximum value for each item = 7.

In order to see if there was significant difference between the + and – complex conditions, likelihood ratio tests were conducted on the responses to the reported mental efforts. Null models included only random effects, whereas increased models contained Complexity as a fixed effect. The results indicated significant difference between the null models and the increased models, Text 1: $\chi^2(1) = 330.03$, $p < .01$, $R^2 = .36$; Text 2: $\chi^2(1) = 70.27$, $p < .01$, $R^2 = .46$. Summaries of the increased mixed effects models revealed that participants in the + complex conditions rated the amount of mental effort significantly greater than those in the – complex conditions for both Text 1: *Estimate* = 1.33, $t = 5.99$ and Text 2: *Estimate* = 2.22, $t = 8.87$. The effect sizes were considered as large. In sum, the mental effort demonstrated that the task manipulation was successful, putting significantly greater cognitive demands.

2. Effects of task complexity and glossing on L2 reading

This section presents results that address RQ (1) and RQ (3), which concern the effects of task complexity and glossing on L2 reading comprehension. More specifically, the following research questions are addressed:

RQ (1) To what extent do the cognitive demands of second language reading tasks affect reading comprehension?

RQ (3) To what extent does glossing affect second language reading comprehension?

Descriptive statistics for the reading comprehension scores of each group are displayed in Table 47. Reading comprehension scores on Text 2 were slightly higher than those on Text 1. A cursory glimpse at the table also reveals that participants in the glossed condition performed better on average than those in the unglossed condition, whereas those in the + complex conditions overall performed worse than those in the – complex conditions.

Table 47. Descriptive statistics for reading comprehension scores

		Text 1			Text 2		
Group	*N*	Mean	*SD*	*SE*	Mean	*SD*	*SE*
[– C, – G]	22	5.82	2.04	.44	6.32	1.59	.34
[– C, + G]	22	6.36	1.43	.31	6.60	1.99	.43
[+ C, – G]	22	5.64	2.46	.53	5.73	1.86	.40
[+ C, + G]	22	5.96	2.17	.46	5.82	1.50	.32
Total	88	5.94	2.04	.22	6.11	1.75	.19

Note. Maximum score = 10.

In order to examine whether task complexity and glossing had a significant impact on L2 reading comprehension scores, linear mixed-effects models were constructed with R. Null models contained random effects (i.e., Subject and Item) only, and each of the fixed effects (i.e., Complexity and Glossing) was entered, one by one, and compared to the null models with a likelihood ratio test using χ^2 statistics. As summarized in Table 48, neither task complexity nor glossing had significant effects on reading comprehension scores. In addition, interaction of Complexity and Glossing had

no significant influence on the reading comprehension scores, $\chi^2(1) = .01$, $p = .92$, $R^2 < .01$.

Table 48. Summary of likelihood ratio tests for predictors on reading comprehension scores

		χ^2	*df*	*p*	R^2
Text 1	Complexity	.87	1	.35	.00
	Glossing	.68	1	.41	.00
Text 2	Complexity	.34	1	$.07^{+}$	.01
	Glossing	.24	1	.63	.00

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

3. Effects of task complexity and glossing on L2 development

This section summarizes the results that address RQ (2) and RQ (4), which concern the effects of task complexity and glossing on development in knowledge of the target constructions. Results for learning unaccusative verbs are presented first, followed by those for target pseudo-words.

RQ (2) To what extent do the cognitive demands of second language reading tasks affect development in the knowledge of target language constructions?

RQ (4) To what extent does glossing affect development in the knowledge of target language constructions?

3.1. Effects of task complexity and glossing on unaccusative verbs

Table 49 presents descriptive statistics for the GJT scores by group. An examination of the table reveals higher mean gain scores in the + complex conditions for the target verbs, but higher mean gain scores in the glossed conditions for novel verbs in the delayed posttests.

To examine whether the participants gained knowledge of the target unaccusative verbs, likelihood ratio tests were conducted comparing null models with random effects (Subject and Item) only and increased models additionally including Time as a fixed effect. The results showed that the participants gained significant GJT gain scores in both the immediate: $\chi^2(1) = 16.74$, $p < .01$, $C = .78$, and the delayed posttest: $\chi^2(1) =$

22.81, $p < .01$, $C = .78$. Then, to these increased models, each of the fixed effects, Complexity and Glossing, was added one by one, and tested in order to investigate whether this inclusion significantly improved the model fit. As summarized in Table 50, Complexity was shown to improve the null model significantly for immediate GJT gain scores.

Table 49. Descriptive statistics for GJT scores (accuracy rate)

Group	Test	*N*	Target verbs			Novel verbs		
			Mean	Mean gain	*SD*	Mean	Mean gain	*SD*
[– C, – G]	Pretest	22	14.64 (48.8%)		3.40	9.77 (61.1%)		1.88
	Immediate posttest	22	14.41 (48.0%)	–.23	4.68	9.41 (58.8%)	–.36	2.68
	Delayed posttest	22	15.41 (51.4%)	.77	4.96	9.73 (60.8%)	–.05	2.59
[– C, + G]	Pretest	22	15.86 (52.9%)		3.21	10.05 (62.8%)		2.21
	Immediate posttest	22	15.41 (51.4%)	1.14	4.77	10.73 (67.1%)	.68	2.90
	Delayed posttest	22	17.46 (58.2%)	1.59	5.59	11.23 (70.2%)	1.18	3.16
[+ C, – G]	Pretest	22	15.68 (52.3%)		3.14	10.59 (66.2%)		2.44
	Immediate posttest	22	17.91 (59.7%)	2.23	5.13	10.55 (65.9%)	–.05	3.02
	Delayed posttest	22	18.41 (61.4%)	2.73	5.50	11.73 (73.3%)	1.14	2.68
[+ C, + G]	Pretest	22	15.18 (50.6%)		3.40	10.14 (63.4%)		2.59
	Immediate posttest	22	18.09 (60.3%)	2.91	5.26	10.91 (68.2%)	.77	2.88
	Delayed posttest	22	17.36 (57.9%)	2.18	5.28	10.77 (67.3%)	.64	2.62

Note. Maximum score for: target GJT items = 30, novel GJT items = 16.

Table 50. Summary of likelihood ratio tests for predictors on GJT gain scores for target verbs

		χ^2	*df*	*p*	*C*
Pretest ~ immediate posttest	Complexity	11.30	2	< .01**	.81
	Glossing	1.78	2	.41	.81
Pretest ~ delayed posttest	Complexity	2.18	2	.34	.81
	Glossing	.04	2	.98	.82

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

The results from the logit mixed-effects models constructed with Complexity as a fixed effect are presented in Table 51. The results showed that Time played a significant role in the model, indicating that learning of the target verbs occurred in the immediate posttest. Also, Complexity was shown to make a significant contribution to

this learning, Time*Complexity: $z = 2.47$, $p = .01$. The *C* index of concordance was .78, which signified a good model fit for the data. In short, participants in this study gained knowledge about the target verbs as manifested in their immediate and delayed posttest scores, and those in the + complex conditions achieved significantly greater GJT gain scores than those in the – complex conditions. Lastly, it was confirmed that interaction of Complexity and Glossing did not affect GJT gain scores, $\chi^2(1) = 1.03$, $p = .60$, $R^2 < .01$.

Table 51. Summary of a mixed-effects model for Time and Complexity on GJT gain scores for target verbs

		Fixed effects				Random effects	
						by Subject	by Item
		Estimate	*SE*	*z*	*p*	*SD*	*SD*
Pretest ~	Intercept	-.23	.17	-1.35	.18	.03	.77
immediate	Time	.29	.09	3.10	< .01**	.28	.29
posttest	Complexity	-.32	.20	-1.64	.10	.06	.12
	Time*Complexity	.38	.15	2.47	.01*	.55	.11

*Formula: GJTtarget ~ Time*Complexity + (Time*Complexity|Subject) + (Time*Complexity|Item); C = .78.*

Note. Pre: Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

Next, another series of likelihood ratio tests were conducted to identify whether the participants' GJT scores for the novel unaccusative verbs increased over the repeated measurements. The results revealed that the participants obtained significant gain scores in the delayed posttest: $\chi^2(1) = 7.79$, $p < .01$, $C = .82$, but not in the immediate posttest: $\chi^2(1) = 1.06$, $p = .30$, $C = .83$. Then, Complexity and Glossing was added one by one to the increased model with the delayed GJT data and tested whether this inclusion had significant influence on the model fit. As summarized in Table 52, significance was not found, indicating neither Complexity nor Glossing had effects on novel GJT scores. Again, interaction of Complexity and Glossing was found to have no significant influence on the novel GJT gain scores, $\chi^2(1) = 5.13$, $p = .08$, $R^2 = .01$.

Table 52. Summary of likelihood ratio tests for predictors on GJT gain scores for novel verbs

		χ^2	*df*	*p*	*C*
Pretest ~ delayed	Complexity	3.43	2	.18	.82
posttest	Glossing	.45	2	.80	.82

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

3.2. Effects of task complexity and glossing on recognition of pseudo-words

Table 53 presents descriptive statistics for vocabulary recognition scores by group. The mean scores on the form recognition test were higher than those on the meaning recognition test. Also, mean form recognition scores from a delayed posttest were higher than those from an immediate posttest, whereas mean meaning recognition scores on the delayed posttest were mostly lower than those on the immediate posttest except for the [- complex, + glossing] condition.

Table 53. Descriptive statistics for vocabulary recognition scores

			Form		Meaning	
Group	Test	*N*	Mean	*SD*	Mean	*SD*
[– C, – G]	Immediate posttest	22	5.68	2.01	1.46	1.14
	Delayed posttest	22	5.86	2.57	1.32	1.21
[– C, + G]	Immediate posttest	22	4.68	2.10	2.05	1.53
	Delayed posttest	22	5.46	2.18	2.27	1.52
[+ C, – G]	Immediate posttest	22	5.50	2.02	1.50	1.68
	Delayed posttest	22	6.09	2.25	.96	1.00
[+ C, + G]	Immediate posttest	22	3.91	2.07	2.14	1.73
	Delayed posttest	22	5.14	2.23	1.64	1.29

Note. Maximum score for: form recognition = 10, meaning recognition = 10.

To identify fixed effects that improved the null models to a significant degree, repeated likelihood ratio tests were conducted using χ^2 statistics. The dependent variables included scores in the immediate and delayed posttests. Again, the null models only included Subject and Item as random effects, and Complexity and Glossing were added, one by one, as fixed effects and tested against the null models. As shown in Table 54, Glossing emerged as a significant factor that improved the null model in an immediate posttest.

Table 54. Summary of likelihood ratio tests for predictors on vocabulary form recognition scores

		χ^2	*df*	*p*	*C*
Immediate posttest	Complexity	1.41	1	.24	.76
	Glossing	5.36	1	.02*	.76
Delayed posttest	Complexity	.10	1	.76	.79
	Glossing	1.43	1	.23	.79

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

With Glossing, a logit mixed-effects model was constructed in maximal random effects structures. As Table 55 shows, Glossing was shown to have a significantly negative influence on vocabulary form recognition scores in an immediate posttest. The *C* index of concordance was .77, which indicated a small effect size. In short, participants in the glossed conditions were less competent at recognizing target word forms in an immediate posttest, compared to those in the unglossed conditions. Interaction of Complexity and Glossing did not influence vocabulary form recognition scores, $\chi^2(1) = .18$, $p = .67$, $R^2 = .01$.

Table 55. Summary of a mixed-effects model for Glossing on immediate vocabulary form recognition scores

		Fixed effects				Random effects	
						by Subject	by Item
		Estimate	*SE*	*z*	*p*	*SD*	*SD*
Immediate posttest	Intercept	.01	.14	.04	.97	.65	.29
	Glossing	-.50	.24	-2.13	.03*	–	.40

Formula: VF ~ Glossing + (1| Subject) + (Glossing| Item); C = .77.

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

Finally, the effects of task complexity and glossing on vocabulary meaning recognition were explored, beginning with another series of likelihood ratio tests using χ^2 statistics. As done previously, the null models included Subject and Item as random effects. The fixed effects, i.e., Complexity and Glossing, were added to the null model, one by one, and examined to see if their inclusion improved the null model to a significant extent. As summarized in Table 56, both Complexity and Glossing were shown to promote the fit of the null model significantly in a delayed posttest.

Table 56. Summary of likelihood ratio tests for predictors on vocabulary meaning recognition scores

		χ^2	*df*	*p*	*C*
Immediate posttest	Complexity	.03	1	.86	.83
	Glossing	3.37	1	$.07^{+}$	.82
Delayed posttest	Complexity	4.71	1	.03*	.79
	Glossing	9.26	1	< .01**	.79

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

Best-fit models were produced starting with a maximal random effects structure with Complexity and Glossing, and random slopes were removed stepwisely until the

models converged. As displayed in Table 57, Glossing seemed to facilitate recognition of target word meanings in a delayed posttest. Additionally, task complexity emerged as a significant factor, negatively affecting meaning recognition. The *C* index of concordance indicated a small effect size (*C* = .78). No collinearity issue between the fixed effects (Complexity and Glossing) was found (condition number = 1.55). In sum, participants who read the glossed texts were better at recognizing word meanings than those who read unglossed texts in a delayed posttest. Also, participants who read texts under + complex conditions had lower scores on a vocabulary meaning recognition test in a delayed posttest than those under the – complex conditions. Interaction of Complexity and Glossing did not have a significant influence on vocabulary recognition scores, $\chi^2(1) = .08$, $p = .78$, $R^2 = .02$.

Table 57. Summary of a mixed-effects model for Complexity and Glossing on vocabulary meaning recognition scores

		Fixed effects				Random effects	
						by Subject	by Item
		Estimate	*SE*	*z*	*p*	*SD*	*SD*
Delayed posttest	Intercept	-2.03	.29	-7.02	< .01**	.04	.80
	Complexity	-.52	.22	-2.36	.02*	.43	.01
	Glossing	.85	.27	3.15	<.01**	.38	.28
	COM*GL	.16	.44	.36	.72	1.00	.14

*Formula: VM~ Complexity*Glossing + (1| Subject) + (Glossing| Item); C = .78.*

Note. COM = Complexity, GL = Glossing; significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

3.3. Interim summary

In sum, in this study, increased task complexity promoted development in the knowledge of target unaccusative verbs in an immediate posttest. Task complexity, however, had significant negative effects on vocabulary meaning recognition in a delayed posttest. In addition, while Glossing facilitated vocabulary meaning recognition in a delayed posttest, it had deteriorating effects on form recognition in an immediate posttest. The following section reports on the nature of knowledge acquired through analysis of various measures of implicit and explicit knowledge (Rebuschat, 2013).

4. Source and solidity of learned knowledge

To explore the source and solidity of acquired knowledge, reaction times, binary confidence ratings and subjective source attributions were analysed further.

4.1. Reaction times for grammaticality judgment tests

As presented in Table 58, the average reaction times taken to respond to the GJT decreased overall for both grammatical and ungrammatical sentences, as compared to those in a pretest. Paired-sample *t*-tests revealed that it took longer overall for participants to respond to ungrammatical sentences than grammatical ones, although this trend narrowly missed significance in a pretest (pretest: $t(87) = -1.98$, $p = .05$, 95% CI [-560.96, 1.17], immediate posttest: $t(87) = -2.04$, $p = .05$, 95% CI [-545.10, -6.81], delayed posttest, $t(87) = -2.48$, $p = .02$, 95% CI [-503.02, -55.67]). Cohen's *d*s ranged from .12 to .16, which were evaluated overall as very small.

Table 58. Average reaction time for GJT (milliseconds)

			Grammatical		Ungrammatical	
Group	Test	N	Mean	*SD*	Mean	*SD*
[– C, – G]	Pretest	22	7,641	1,900	8,013	2,567
	Immediate posttest	22	6,628	1,865	6,901	1,684
	Delayed posttest	22	5,627	1,602	5,887	1,738
[– C, + G]	Pretest	22	7,359	1,815	7,596	2,161
	Immediate posttest	22	6,260	1,859	6,266	1,713
	Delayed posttest	22	5,400	1,562	5,677	2,131
[+ C, – G]	Pretest	22	7,620	2,700	7,906	3,233
	Immediate posttest	22	5,888	1,974	6,381	2,198
	Delayed posttest	22	5,405	1,674	5,771	1,897
[+ C, + G]	Pretest	22	6,812	1,938	7,037	1,877
	Immediate posttest	22	5,518	1,949	5,850	1,827
	Delayed posttest	22	5,205	1,765	5,418	1,843

Next, it was investigated whether reaction times to the GJT target items decreased over the repeated measurements using likelihood ratio tests. The null models contained only the random effects (Subject and Item), and Time was added to the null models. The results revealed that reaction times decreased significantly from the pretest to the immediate: $\chi^2(1) = 332.88$, $p < .01$, $R^2 = .04$, and the delayed posttest: $\chi^2(1) = 712.5$, $p < .01$, $R^2 = .09$. Next, Complexity and Glossing was added one by one to the models

that included Time as an existing fixed effect and tested against them in order to examine whether the inclusion improved the model fit to a significant extent. As displayed in Table 59, Complexity emerged as a significant factor in the reaction time data for the immediate posttest. The summary of post-hoc mixed-effects modelling confirmed that, as shown in Table 60, Complexity had a significant influence on the decreasing trend in the reaction time data in the immediate posttest.

For the reaction time data for the novel GJT items, another series of likelihood ratio tests were performed. Time, again, emerged as a significant factor in the immediate: $\chi^2(1) = 143.35$, $p < .01$, $R^2 = .03$, and the delayed posttest: $\chi^2(1) = 308.64$, $p < .01$, $R^2 = .07$, indicating that reaction times taken to the novel GJT items decreased significantly over the repeated measurements. Yet, as shown in Table 59, neither Complexity nor Glossing emerged as a significant predictor.

Table 59. Summary of likelihood ratio tests for predictors on RTs to GJT tests

Target		χ^2	*df*	*p*	R^2
Pretest ~ immediate posttest	Complexity	16.34	2	< .01**	.05
	Glossing	1.68	2	.43	.04
Pretest ~ delayed posttest	Complexity	.67	2	.71	.09
	Glossing	1.66	2	.44	.09
Novel					
Pretest ~ immediate posttest	Complexity	2.07	2	.35	.04
	Glossing	1.77	2	.41	.04
Pretest ~ delayed posttest	Complexity	2.84	2	.24	.08
	Glossing	1.45	2	.48	.08

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

Table 60. Summary of mixed-effects models for interaction among Time, Complexity and Glossing on reaction times to GJT items

	Fixed effects			Random effects	
				by Subject	by Item
	Estimate	*SE*	*t*	*SD*	*SD*
Intercept	9.09	.04	203.51°	.18	.17
Time	-.21	.01	-18.57°	–	–
Complexity	.06	.06	.93	.35	–
Time*Complexity	-.09	.02	-3.85°	–	–

*Formula: log(RT) ~Time*Complexity + (Time*Complexity | Subject) + (1 | Item); $R^2 = .03$.*

Note. RT = Reaction times; Significance: °| *t* | > 2.0.

4.2. Reaction times for vocabulary recognition tests

Table 61 displays average reaction times in vocabulary recognition tests; reaction

times appeared to decrease overall from immediate to delayed posttests. Paired-sample *t*-tests also revealed that meaning recognition took significantly longer than form recognition in both immediate: $t(45) = -9.81$, $p < .01$, 95% CI [-1914.00, -1302.63] and delayed posttests: $t(51) = -2.80$, $p < .01$, 95% CI [-1656.68, -1088.56]. Cohen's *d*s were 1.71 and 1.38, which indicated large effect sizes.

Table 61. Average reaction time for vocabulary recognition tests (milliseconds)

			Form			Meaning	
Group	Test	*N*	Mean	*SD*	*N*	Mean	*SD*
[– C, – G]	Immediate posttest	22	1,977	453	22	4,040	643
	Delayed posttest	22	1,820	635	22	2,972	1,224
[– C, + G]	Immediate posttest	22	2,118	643	22	3,317	1,039
	Delayed posttest	22	2,116	987	22	3,190	1,225
[+ C, – G]	Immediate posttest	22	2,006	1,026	22	3,551	1,448
	Delayed posttest	22	1,816	641	22	3,577	1,261
[+ C, + G]	Immediate posttest	22	2,168	860	22	3,704	1,130
	Delayed posttest	22	1,487	502	22	3,034	1,150

Note. Missing values were excluded analysis by analysis.

Again, to examine whether Complexity and Glossing had any effects on the decreasing trend in the reaction times to vocabulary recognition tests, likelihood ratio tests were conducted using χ^2 statistics. Null models were constructed only with random effects on either form or meaning recognition test scores, and each of the fixed effects, i.e., Complexity and Glossing, was entered and tested for its ability to improve the fit of the null models. As displayed in Table 62, it was found that neither Complexity nor Glossing had a significant influence on the reaction times to the vocabulary form and meaning recognition tests.

Table 62. Summary of likelihood ratio tests for predictors on RTs to vocabulary recognition tests

Form		χ^2	*df*	*p*	R^2
Immediate ~ delayed	Complexity	2.82	1	$.09^{+}$	.00
Posttests	Glossing	.38	1	.54	.00
Meaning					
Immediate ~ delayed	Complexity	1.19	1	.28	.00
Posttests	Glossing	1.16	1	.28	.00

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

4.3. Confidence ratings for grammaticality judgment tests

Prior to investigating participants' confidence ratings for the GJT, gain scores in the immediate and delayed posttests were analysed using one-sample *t*-tests against zero. As presented in Table 63, gain scores were significantly greater than zero for the immediate target, delayed target and delayed novel items in the [+ complex, – glossing] condition. Gain scores were also found to be significant for the immediate target and delayed target items in the [+ complex, + glossing] condition. That is, development seemed to occur in the complex conditions, especially for target verbs.

Table 63. Significance of gain scores for GJT and *d'* values by group

		Mean	*t*	*d'*	*t*
[– C, – G]	Immediate_target	-.23	-.26		
	Delayed_target	.77	.81		
	Immediate_novel	.36	-.83		
	Delayed_novel	.05	-.11		
[– C, + G]	Immediate_target	1.14	1.47		
	Delayed_target	1.59	1.89		
	Immediate_novel	.68	1.09		
	Delayed_novel	1.18	1.99		
[+ C, – G]	Immediate_target	2.23	2.66*	.23	2.20*
	Delayed_target	2.73	2.82*	.33	2.02^{+}
	Immediate_novel	-.05	-.09		
	Delayed_novel	1.14	2.76*	.78	2.52*
[+ C, + G]	Immediate_target	2.91	3.91**	.10	.50
	Delayed_target	2.18	2.81*	.13	.68
	Immediate_novel	.77	1.73		
	Delayed_novel	.64	1.25		

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

Next, for each GJT datum where significant gains were found, sensitivity index *d'* was calculated using the technique developed by Kunimoto et al. (2001). As mentioned earlier, a *d'* of, or below, zero indicates no awareness, whereas a positive *d'* signifies awareness. As can be seen in the table, *d'* was significantly higher than zero for immediate target items and for delayed novel items in the [+ complex, – glossing] condition, while narrowly missing significance for delayed target items. In the [+ complex, + glossing] condition, *d'* was not significantly above zero. In other words, increased complexity might have contributed to the participants' increased level of

confidence in their correct responses, whereas participants were less confident in their responses when complexity was combined with glossing.

4.4. Confidence ratings for vocabulary recognition tests

Gain scores in the vocabulary recognition tests were significantly greater than zero for all experimental conditions, indicating a significant amount of development. As seen in Table 64, *d'* values for form recognition were significantly greater than zero in the [– complex, – glossing] condition, Immediate posttest: $t(21) = 2.56$, $p = .02$, 95% CI [.16, 1.57], Delayed posttest: $t(21) = 2.17$, $p = .04$, 95% CI [.03, 1.38]. Cohen's *d*s were .77 and .67, which indicated large and medium effect sizes. In the rest of the cases, *d'* values were not significantly above zero.

Table 64. Significance of gain scores and *d'* values for vocabulary form recognition

Group	Test	Mean	*t*	*d'*	*t*
[– C, – G]	Immediate posttest	5.68	13.26**	.87	2.56*
	Delayed posttest	5.86	10.71**	.70	2.17*
[– C, + G]	Immediate posttest	4.77**	10.45**	.04	.12
	Delayed posttest	5.46**	11.76**	-.26	-.80
[+ C, – G]	Immediate posttest	5.46**	12.79**	.48	1.11
	Delayed posttest	5.91**	12.73**	.38	1.26
[+ C, + G]	Immediate posttest	3.91**	8.87**	.48	1.18
	Delayed posttest	5.14**	10.80**	.17	.48

Note. Significantly above zero: * $p < .05$, ** $p < .01$.

As Table 65 displays, *d*' for meaning recognition was not significantly greater than zero in all cases. It seems notable that *d*' values for meaning recognition were all below zero, whereas those for form recognition were mostly above zero. That is, overall, participants appeared to be more confident in their correct responses to vocabulary form recognition tests compared to meaning recognition tests.

Table 65. Significance of gain scores and *d'* values for vocabulary meaning recognition

Group	Test	Mean	*t*	*d'*	*t*
[– C, – G]	Immediate posttest	1.46	5.97**	-.56	-1.80
	Delayed posttest	1.41	5.11**	-1.16	-3.67
[– C, + G]	Immediate posttest	2.00	6.28**	-.30	.91
	Delayed posttest	2.27	7.02**	-1.21	.43
[+ C, – G]	Immediate posttest	1.50	4.18**	-.62	-2.36
	Delayed posttest	0.86	4.48**	-.36	-1.03
[+ C, + G]	Immediate posttest	2.14	5.80**	-.95	-3.37
	Delayed posttest	1.64	5.94**	-.80	-2.24

Note. Significantly above zero: * $p < .05$, ** $p < .01$.

4.5. Source attribution for grammaticality judgment tests

An analysis of source attribution was conducted for cases where GJT gain scores were significantly larger than zero. The proportions of source attributions revealed that participants' grammaticality judgments were mostly based on intuition and rules (see Table 66). The accuracy results provided similar findings to those of confidence ratings. That is, participants assigned to the [+ complex, – glossing] condition tended to respond correctly when their judgments were based on memory and rules, whereas those in the [+ complex, + glossing] condition responded correctly when their judgments were based on guesswork and intuition. This may imply that participants in the [+ complex, – glossing] condition were, overall, aware of the rules when responding to GJT items, whereas those in the [+ complex, + glossing] were relatively unaware.

Table 66. Mean proportions and mean accuracy rates across source attribution

			Guess	Intuition	Memory	Rule
[+ C, – G]	Immediate_target	Proportion	.14	.38	.19	.29
		Accuracy	.60	.59	.53	.60
	Delayed_target	Proportion	.12	.36	.17	.36
		Accuracy	.54	.54	.61*	.66*
	Delayed_novel	Proportion	.09	.30	.21	.41
		Accuracy	.69	.73*	.77**	.69*
[+ C, + G]	Immeidate_target	Proportion	.08	.37	.23	.35
		Accuracy	.68*	.58	.57	.58
	Delayed_target	Proportion	.12	.38	.18	.49
		Accuracy	.49	.60*	.60	.55

Note. Significant above chance (.50): $^{+}p < .1$, *$p < .05$, **$p < .01$.

4.6. Interim summary

The reaction time data revealed, overall, a similar pattern to the findings in Study 1. That is, it took significantly longer for participants to respond to ungrammatical sentences than to grammatical ones, and to a vocabulary meaning recognition test than to a form recognition test. Again, neither task complexity nor glossing resulted in changes to the reaction time data for the GJT and vocabulary recognition tests. Additionally, the binary confidence ratings demonstrated that participants were in general more confident about the accuracy of their responses in vocabulary form recognition tests, but uncertain about meaning recognition tests. Yet, when it comes to the GJT scores, unlike Study 1 where participants in the + complex condition were more confident when they read glossed texts, in Study 2 the participants in the + complex condition reported that they were more confident when they read unglossed texts.

5. WMC as a moderator of L2 reading and L2 learning

This section displays the results for the moderating effects of working memory capacity. More specifically, the following research questions were addressed:

RQ (5) To what extent does working memory capacity moderate the effects of the cognitive demands of second language reading tasks on second language reading comprehension?

RQ (6) To what extent does working memory capacity moderate the effects of the cognitive demands of second language reading tasks on development in the knowledge of target language constructions?

RQ (7) To what extent does working memory capacity moderate the effects of glossing on second language reading comprehension?

RQ (8) To what extent does working memory capacity moderate the effects of glossing on development in the knowledge of target language constructions?

To examine whether working memory measures tapped into related constructs, Pearson's correlation coefficient was calculated for digit span scores, nonword span

scores, backward digit span scores and operation span scores. As summarized in Table 67, digit span scores, nonword span scores and backward digit span scores correlated significantly with each other. For operation span scores, however, significant correlation was found only with backward digit span scores. That is, digit span scores, nonword span scores and backward digit span scores appeared to measure common constructs, most probably aspects of phonological short-term memory capacity. Also, backward digit span scores and operation span scores appeared to estimate partially overlapping constructs, presumably components of complex working memory capacity. The significant correlations were, however, small in all cases.

Table 67. Correlations among working memory capacity indices

		DS	NWS	BDS	OSPAN
DS	Coefficient	1	.46**	.44**	.17
	Significance		.00	.00	.13
NWS	Coefficient		1	.35**	.17
	Significance			.00	.11
BDS	Coefficient			1	.39**
	Significance				.00
OSPAN	Coefficient				1
	Significance				

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

To test whether working memory moderated the effects of Complexity and Glossing on reading comprehension scores, likelihood ratio tests were conducted using χ^2 statistics. Null models included Complexity and Glossing as fixed effects and Subject and Item as random effects. To these null models, each of digit span scores, nonword span scores, backward digit span scores and operation span scores was entered, one by one, to see if their inclusion improved the model fit significantly. As shown in Table 68, digit span scores and backward digit span scores improved the null models for Text 1, and operation span scores were revealed as a significant factor in Text 2.

Table 68. Summary of likelihood ratio tests for interaction among WMC, Complexity and Glossing on reading comprehension scores

		χ^2	*df*	*p*	R^2
Text 1	DS	15.49	4	< .01**	.03
	NWS	8.16	4	.09^{+}	.02
	BDS	13.80	4	< .01**	.02
	OSPAN	9.13	4	.06^{+}	.02
Text 2	DS	3.29	4	.51	.01
	NWS	2.99	4	.56	.01
	BDS	3.11	4	.54	.01
	OSPAN	21.88	4	< .01**	.04

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

Table 69 presents the results from linear mixed-effects models that included the working memory capacity indices shown to improve the null models. Whenever the models failed to converge, random parameters were dropped from the one that accounted for the least variance to the next until convergence was reached. Significant interaction emerged between digit span scores and Glossing in Text 1, among backward digit span scores, Complexity and Glossing in Text 1, and among operation span scores, Complexity and Glossing in Text 2. R^2s of the models ranged from .02 to .04, indicating very small effect sizes.

Table 69. Summary of mixed-effects models for interaction among WMC, Complexity and Glossing on reading comprehension scores

		Fixed effects			Random effects	
					by Subject	by Item
		Estimate	*SE*	*t*	*SD*	*SD*
Text 1	Intercept	.54	.24	2.27°	.13	.26
	COM*DS	-.06	.05	-1.22	–	–
	GL*DS	-.15	.05	-3.17°	–	–
	COM*GL*DS	-.15	.10	-1.53	–	–
*Formula: RC ~ Complexity * Glossing * DS + (1 | Subject) + (1 | Item); R^2 = .03.*						
	Intercept	.65	.14	4.57°	.13	.26
	COM*BDS	-.03	.03	-1.24	–	–
	GL*BDS	-.05	.03	-1.70	–	–
	COM*GL*BDS	-.20	.06	-3.30°	–	–
*Formula: RC ~ Complexity * Glossing * BDS + (1 | Subject) + (1 | Item); R^2 = .02.*						
Text 2	Intercept	-.01	.25	-.04	.09	.20
	COM*OSPAN	-.01	.01	-1.46	–	–
	GL*OSPAN	.00	.01	.50	–	–
	COM*GL*OSPAN	-.05	.02	-3.59°	–	–
*Formula: RC ~ Complexity * Glossing * OSPAN + (1 | Subject) + (1 | Item); R^2 = .04.*						

Note. COM = Complexity, GL = Glossing; significance: °| *t* | > 2.0.

Next, post hoc mixed-effects models were constructed to examine the differential contribution made by working memory to reading comprehension scores across experimental conditions. As shown in Table 70, significant effects were found for digit span scores in the glossed conditions and backward digit span scores in the [+ complex, + glossing] condition for Text 1. Significance was also found for operation span scores in the [– complex, + glossing] condition for Text 2. In other words, for Text 1, when assigned to the glossed conditions, participants with higher digit span scores were worse at answering reading comprehension questions than those with lower digit span scores. Also, in the [+ complex, + glossing] condition, participants with higher backward digit span scores read Text 1 better than those with lower backward digit span scores. For Text 2, participants with higher operation span scores performed better than those with lower operation span scores in the [– complex, + glossing] condition. R^2s ranged from .02 to .09, which were considered as small effect sizes.

Table 70. Summary of post-hoc mixed-effects models for interaction among WMC, Complexity and Glossing on reading comprehension scores

		Fixed effects			Random effects	
					by-Subject	by-Item
		Estimate	*SE*	*t*	*SD*	*SD*
Text 1	Intercept	-.08	.37	-.22	.13	.23
- GL	DS	.08	.04	1.82	.03	.06
	Formula: RC ~ DS + (DS|Subject) + (DS|Item); R^2 = .01.					
+ GL	Intercept	1.34	.28	4.86°	.61	.25
	DS	-.07	.03	-2.45°	.08	.00
	Formula: RC ~ DS + (DS|Subject) + (DS|Item); R^2 = .02.					
- GL, - COM	Intercept	.87	.23	3.83°	.51	.05
	BDS	-.03	.03	-.80	.07	.04
	Formula: RC ~ BDS + (BDS|Subject) + (BDS|Item); R^2 = .00.					
+ GL, - COM	Intercept	.40	.18	2.29°	.20	.13
	BDS	.04	.02	1.82	.02	.01
	Formula: RC ~ BDS + (BDS|Subject) + (BDS|Item); R^2 = .02.					
- GL, + COM	Intercept	.16	.37	.44	.66	.32
	BDS	.06	.05	1.30	.06	.05
	Formula: RC ~ BDS + (BDS|Subject) + (BDS|Item); R^2 = .02.					
+ GL, + COM	Intercept	1.31	.23	5.63°	.02	.22
	BDS	-.10	.03	-2.86°	.02	.01
	Formula: RC ~ BDS + (BDS|Subject) + (BDS|Item); R^2 = .05.					
Text 2	Intercept	.42	.71	.58	.74	1.40
- GL, - COM	OSPAN	.01	.01	.43	.01	.02
	Formula: RC ~ OSPAN + (OSPAN|Subject) + (OSPAN|Item); R^2 = .00.					
+ GL, - COM	Intercept	-1.28	.66	-1.93	.09	1.44
	OSPAN	.03	.01	3.10°	–	.00
	Formula: RC ~ OSPAN + (1|Subject) + (OSPAN|Item); R^2 = .09.					
- GL, + COM	Intercept	-.44	.55	-.80	.11	.57
	OSPAN	.02	.01	1.93	–	.01
	Formula: RC ~ OSPAN + (1|Subject) + (OSPAN|Item); R^2 = .03.					
+ GL, + COM	Intercept	1.03	.36	2.86°	.04	.18
	OSPAN	-.01	.01	-1.09	–	–
	Formula: RC ~ OSPAN + (1|Subject) + (1|Item); R^2 = .01.					

Note. COM = Complexity, GL = Glossing; significance: °| *t* | > 2.0.

The role of working memory capacity as a moderator of the effects of Complexity and Glossing was also investigated in relation to GJT gain scores. Likelihood ratio tests were conducted comparing reduced models that included Time, Complexity and Glossing as existing fixed effects and Subject and Item as random effects with increased models that contained each of the working memory capacity indices as additional fixed effects. As summarized in Table 71, digit span scores, backward digit span scores and operation span scores were found to improve the reduced models for target GJT gain

scores significantly. Accordingly, mixed-effects models were constructed with these working memory capacity measures.

Table 71. Summary of likelihood ratio tests for interaction among WMC, Complexity and Glossing on GJT gain scores

Target		χ^2	*df*	*p*	*C*
Pretest ~ immediate posttest	DS	21.132	8	< .01**	.77
	NWS	2.226	8	.97	.77
	BDS	8.867	8	.35	.77
	OSPAN	16.070	8	< .01**	.77
Pretest ~ delayed posttest	DS	22.64	8	< .01**	.78
	NWS	2.67	8	.95	.78
	BDS	16.074	8	.04*	.78
	OSPAN	5.460	8	.71	.78
Novel					
Pretest ~ immediate posttest	DS	6.940	8	.54	.83
	NWS	3.601	8	.98	.83
	BDS	10.704	8	.22	.83
	OSPAN	5.009	8	.76	.83
Pretest ~ delayed posttest	DS	8.472	8	.39	.82
	NWS	6.115	8	.63	.82
	BDS	15.458	8	.05^{+}	.82
	OSPAN	10.560	8	.23	.82

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

Mixed-effects models included interaction among Complexity, Glossing and Time (i.e., changes in GJT scores from pretest scores) and each of the working memory capacity indices that made a significant improvement to the null models in the likelihood ratio tests. Best-fit models were sought through backwards elimination of random parameters that accounted for the least variance in the data stepwisely from the maximal random effects models. In all cases, models converged only when the random parameters were removed, except for random intercepts. As displayed in Table 72, digit span scores moderated the interaction effect among Time, Complexity and Glossing in immediate gain scores. *C* index of concordance was .75, which indicated a moderate model fit for the data.

Table 72. Summary of mixed-effects models for interaction among WMC, Complexity and Glossing on target GJT gain scores

		Fixed effects				Random effects	
						by Subject	by Item
Target		Estimate	*SE*	*z*	*p*	*SD*	*SD*
Pretest ~	Intercept	.23	1.16	.20	.84	.52	.97
immediate	Time*COM*DS	-.06	.14	-.46	.64	–	–
posttest	Time*GL*DS	-.11	.14	-.82	.41	–	–
	Time*COM*GL*DS	-.60	.27	-2.25	.03*	–	–
*Formula: GJTtarget ~ Time * Complexity * Glossing * DS + (1 | Subject) + (1 | Item); C = .77.*							
	Intercept	1.12	1.46	.76	.45	.56	.97
	Time*COM*OSPAN	-.04	.02	-1.74	.09^{+}	–	–
	Time*GL*OSPAN	.01	.02	.54	.59	–	–
	Time*COM*GL*OSPAN	.01	.05	.25	.81	–	–
*Formula: GJTtarget ~ Time * Complexity * Glossing * OSPAN + (1 | Subject) + (1 | Item); C = .77.*							
Pretest ~	Intercept	.56	.93	.60	.55	.54	.96
delayed	Time*COM*DS	-.02	.07	-.27	.79	–	–
posttest	Time*GL*DS	.11	.07	1.59	.11	–	–
	Time*COM*GL*DS	-.21	.14	-1.50	.13	–	–
*Formula: GJTtarget ~ Time * Complexity * Glossing*DS + (1 | Subject) + (1 | Item); C = .78.*							
	Intercept	.37	.50	.74	.46	.56	.96
	Time*COM*BDS	-.06	.04	-1.38	.17	–	–
	Time*GL*BDS	.08	.04	1.79	.07^{+}	–	–
	Time*COM*GL*BDS	.04	.08	.51	.61	–	–
*Formula: GJTtarget ~ Time * Complexity * Glossing * BDS + (1 | Subject) + (1 | Item); C = .78.*							

Note. COM = Complexity, GL = Glossing; significance: $^{+}p < .1$, $*p < .05$, $**p < .01$.

Next, a series of mixed-effects modelling tests was performed to examine the effects of the relationship between working memory capacity and fixed effects on GJT gain scores for target verbs. As presented in Table 73, digit span scores had significant moderating effects on immediate GJT gain scores in the [+ complex, + glossing] condition. That is, when participants read glossed texts under the + complex conditions, those with higher digit span scores obtained lower GJT gain scores in an immediate posttest than those with lower digit span scores. *C* index of concordance was .79, signifying a moderate model fit for the data.

Table 73. Summary of post-hoc mixed-effects models for interaction among DS, Complexity and Glossing on immediate GJT gain scores for target verbs

		Fixed effects				Random effects	
						by Subject	by Item
		Estimate	*SE*	*z*	*p*	*SD*	*SD*
- GL, - COM	Intercept	1.01	2.18	.46	.64	.18	.60
	Time	2.00	1.57	1.28	.20	.32	.29
	DS	-.11	.23	-.49	.63	–	–
	Time*DS	-.22	.17	-1.29	.20	–	–
*Formula: GJTtarget ~ Time*DS + (Time|Subject) + (Time|Item); C = .79.*							
+ GL, - COM	Intercept	-.11	1.67	-.06	.95	.06	.97
	Time	-.11	1.30	-.09	.93	.42	.38
	DS	.00	.18	.02	.98	–	–
	Time*DS	.03	.14	.23	.82	–	–
*Formula: GJTtarget ~ Time*DS + (Time|Subject) + (Time|Item); C = .79.*							
- GL, + COM	Intercept	2.39	2.48	.96	.34	.20	.49
	Time	.08	2.09	.04	.97	.52	.36
	DS	-.28	.26	-1.09	.28	–	–
	Time*DS	.04	.22	.17	.86	–	–
*Formula: GJTtarget ~ Time*DS + (Time|Subject) + (Time|Item); C = .79.*							
+ GL, + COM	Intercept	-3.55	1.73	-2.06	.04*	.36	.94
	Time	3.48	1.10	3.16	< .01**	.19	.33
	DS	.34	.19	1.80	.07	–	–
	Time*DS	-.33	.12	-2.73	.01*	–	–
*Formula: GJTtarget ~ Time*DS + (Time|Subject) + (Time|Item); C = .79.*							

Note. COM = Complexity, GL = Glossing; significance: $^{+}p < .1$, $*p < .05$, $**p < .01$.

Last but not least, likelihood ratio tests were conducted in order to identify working memory capacity indices that improved reduced models in vocabulary recognition scores. As summarized in Table 74, only operation span scores improved the reduced models in delayed meaning recognition scores, and thus operation span scores were included in post hoc mixed-effects models.

Table 74. Summary of likelihood ratio tests for interaction among WMC, Complexity and Glossing on vocabulary recognition scores

Form		χ^2	*df*	*p*	*C*
Immediate	DS	1.17	8	.88	.76
	NWS	1.64	8	.82	.76
	BDS	.96	8	.97	.76
	OSPAN	1.95	8	.74	.75
Delayed	DS	2.91	8	.57	.79
	NWS	.76	8	.94	.79
	BDS	4.20	8	.38	.79
	OSPAN	9.06	8	.06^{+}	.79
Meaning					
Immediate	DS	2.71	8	.61	.81
	NWS	.75	8	.95	.81
	BDS	1.74	8	.78	.81
	OSPAN	6.75	8	.15	.80
Delayed	DS	3.62	8	.46	.71
	NWS	5.73	8	.22	.77
	BDS	4.92	8	.30	.77
	OSPAN	15.05	8	.01**	.75

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

For vocabulary meaning delayed posttests, operation span scores were additionally entered into the reduced model, and backwards elimination from a maximal random effects structure was conducted in stepwise fashion from the random parameter accounting for the least variance to the next. As Table 75 displays, significant influence was found for the interaction between Complexity and operation span scores in delayed vocabulary meaning recognition scores. Model fit was evaluated as moderate, as manifested in the *C* index of concordance, $C = .75$.

Table 75. Summary of a mixed-effects model for interaction among OSPAN, Complexity and Glossing in delayed vocabulary meaning recognition scores

		Fixed effects				Random effects	
						by Subject	by Item
		Estimate	*SE*	*z*	*p*	*SD*	*SD*
Delayed	Intercept	-.70	1.37	-.51	.61	.14	.77
	COM*OSPAN	.15	.04	3.64	<.01**	–	–
	GL*OSPAN	-.01	.04	-.19	.85	–	–
	COM*GL*OSPAN	-.02	.08	-.20	.85	–	–

*Formula: VM ~ Complexity * Glossing * OSPAN + (1 | Subject) + (1 | Item); C = .75.*

Note. COM = Complexity, GL = Glossing; $^{+}p < .1$, $*p < .05$, $**p < .01$.

In order to determine the differential effects of the interaction between operation span scores and task complexity across the + and – complex conditions, a post hoc mixed-effects model was constructed for delayed vocabulary meaning recognition

scores. As shown in Table 76, operation span emerged as a significant factor for the – complex conditions. That is to say, when performing the – complex version, participants with higher operation span scored did worse on a delayed vocabulary meaning recognition test than those with lower operation span scores. *C* index of concordance was .76, indicating a moderate model fit for the data.

Table 76. Summary of post-hoc mixed-effects models for interaction among OSPAN, Complexity and Glossing in delayed vocabulary meaning recognition scores

		Fixed effects				Random effects	
						by Subject	by Item
		Estimate	*SE*	*z*	*p*	*SD*	*SD*
- COM	Intercept	4.02	1.82	2.20	.03*	.31	.72
	OSPAN	-.09	.03	-3.12	< .01**	–	–
Formula: VM ~ OSPAN + (1|Subject) + (1|Item); C = .76.							
+ COM	Intercept	-5.46	2.06	-2.66	.01*	.49	.69
	OSPAN	.05	.03	1.62	.10	–	–
Formula: VM ~ OSPAN + (1|Subject) + (1|Item); C = .81.							

Note. COM = Complexity, GL = Glossing; significance: $^{+}p < .1$, $*p < .05$, $**p < .01$.

V. Interim Discussion

This study replicated Study 1, addressing the same research questions. To be more specific, it was investigated whether manipulating task features in terms of cognitive demands and glossing reading texts affected learners' reading comprehension and development in their knowledge of target linguistic constructions. As in Study 1, working memory capacity was investigated as a potential moderator factor. The major difference between Studies 1 and 2 lay in the way task complexity was operationalised. In Study 1, disarrangement of texts was conducted on a discourse level by jumbling the order of paragraphs, whereas in Study 2, it was operated on a near sentential level by disorganizing sentences in each paragraph. The results of Study 2 are discussed in depth in the following section, focusing on each of the research questions (for a summary of significant results, see Table 77).

Table 77. Summary of significant results of Study 2

	Dependent variables	Statistical results
Fixed effects		
Complexity	Immediate target GJT gain scores	$z = 2.47$, $p = .01$*, $C = .78$
	Delayed word meaning recognition scores	$z = -2.36$, $p = .02$*, $C = .78$
Glossing	Immediate word form recognition scores	$z = -2.13$, $p = .03$*, $C = .77$
	Delayed word meaning recognition scores	$z = 3.15$, $p < .01$**, $C = .78$
Moderator		
DS	Reading comprehension for Text 1 in [+ GL]	$t = -2.45$°, $R^2 = .02$
	Immediate target GJT gain scores in [+ C, + G]	$z = -2.73$, $p = .01$*, $C = .79$
BDS	Reading comprehension for Text 1 in [+ C, + G]	$t = -2.86$°, $R^2 = .05$
OSPAN	Reading comprehension for Text 2 in [- C, + G]	$t = 3.10$°, $R^2 = .09$
	Delayed word meaning recognition scores in [- COM]	$z = -3.12$, $p < .01$**, $C = .76$

Note. Significance: °| t | > 2.0; $^{+}p < .1$, *$p < .05$, **$p < .01$.

1. Effects of task complexity and glossing on L2 reading comprehension

Again, as in Study 1, both task complexity and glossing failed to have significant effects on reading comprehension. Although scores were slightly higher in the glossed conditions than the unglossed conditions, as well as lower in the + complex conditions than in the – complex conditions, the differences were not statistically significant. Given that only one or two reading comprehension items followed each paragraph, the chances to observe meaningful effects of task complexity and/or glossing on reading comprehension could have been restricted. In addition, it is also noteworthy that the participants were from a highly homogeneous population, placing a limitation on expanding variances in the scores across task conditions. It is equally possible that task manipulation could be detected in the online reading processes, if not in the comprehension outcomes. Indeed, previous research has repeatedly shown that task effects are manifested by process measures, such as think-aloud protocols, but not by comprehension measures in various formats, such as oral summary recall (Horiba, 2000), free written recall (Horiba, 2013), cloze tests, open-ended questions and summary writing (Kobayashi, 2002) and true-or-false comprehension check questions (Yoshimura, 2006).

2. Effects of task complexity on development in the knowledge of target constructions

In Study 2, task manipulation was designed to operate on a more localized level in comparison to Study 1, where tasks were modified on a discourse level by disarranging the order of paragraphs. Sentential-level task manipulation, as intended, seemed effective in increasing the cognitive demands put on the participants, presumably leading them to process the given texts more thoroughly. In addition, there was a 25-minute time limit for task completion, unlike in Study 1 where no time limit was set to collect participants' subjective time estimations. Perhaps due to these changes, unlike in Study 1, participants assigned to the + complex conditions rated the amount of mental effort to complete the tasks significantly greater than those in the – complex conditions. Probably as a result, task complexity was shown to have significant effects on their development in target English unaccusative verbs in an immediate posttest (mean gain scores: .46 in – complex, 2.57 in + complex, $C = .78$). Participants under the + complex conditions were significantly faster in responding to the immediate target GJT items than those under the – complex conditions, which seems to lend further credence to the facilitative effects of increased task complexity on promoting acquisition of the target unaccusative verbs. Although the significant influence did not last for a delayed posttest, the immediate effects of task complexity appear to indicate that the participants had to read each sentence more meticulously and repeatedly, in order to arrange the sentences in a coherent order in Study 2. However, as in Study 1, no transfer of learning to novel unaccusative verbs was found. Given the learnability problem of English unaccusative verbs ascribed to multiple factors such as the presence or absence of an alternating transitive counterpart (Balcom, 1997; Hwang, 1999, 2001), lack of a conceptualizable agent (Ju, 2000), L1 transfer (No & Chung, 2006) and low input frequency (Lee et al., 2008), it seems understandable that a single exposure to each target verb was not

enough for long-term development in the knowledge of correct usage of those verbs and further transfer of learning.

Another difference from Study 1 was that task complexity did not affect vocabulary form recognition scores. A possible explanation of this finding is that the increased level of task complexity compared to that of Study 1, in addition to the 25-minute time limit, might have inhibited participants from noticing the forms of target pseudo-words in both conditions. In other words, as long as reading comprehension was not interrupted, participants might have allocated their mental resources to the text-ordering task and thus not bothered to focus on pseudo-word forms. In addition, although serious consideration was given to selecting target words for the experimental texts, pseudo-words might have not been essential for completing the text-ordering task. If the pseudo-words had been crucial for answering the reading comprehension questions or determining the correct order of the sentences, task complexity could have influenced the processing of word forms, and in turn affected word form recognition scores.

The long-term negative effects of increased task complexity on meaning recognition scores might be explained by the same logic (delayed meaning recognition scores: 1.30 in + complex, 1.80 in – complex, *C* index = .78). That is, the greater cognitive load put on the participants might have resulted in depleted attentional resources, leaving only a slim chance of noticing and inferring the meanings of unknown words. Another viable account of this finding is that jumbled text segments, lacking in coherence, might have resulted in inadequate contextual clues and interrupted inferring the meanings of pseudo-words.

3. Effects of glossing on development in the knowledge of target constructions

As in Study 1, glossing was shown to affect receptive knowledge of the forms and meanings of pseudo-words. More specifically, glossing had a significantly negative influence on vocabulary form recognition (delayed form recognition scores: 5.98 in – glossed, 5.30 in + glossed, *C* index = .77), but a positive impact on meaning recognition (delayed meaning recognition scores: 1.14 in – glossed, 1.95 in + glossed, *C* index = .78). The facilitative effects of glossing on word meaning recognition scores seem to corroborate previous studies that found positive effects for glosses on learning of L2 lexical features (e.g., Chun & Plass, 1996; Hulstjin et al., 1996; Ko, 2012; Martinez-Fernández, 2010; Watanabe, 1997). That is, the meanings provided by glosses seemed to be noticed by participants in the process of reading texts and doing text-arranging tasks.

The negative effects of glossing on word form recognition, however, run counter to the findings from Study 1, where glossing significantly promoted form recognition scores. The contradictory findings might be explained by the overall differences in task conditions between Studies 1 and 2. More specifically, in Study 1 where relatively lower cognitive demands appeared to be imposed on the learners, the participants could have had sufficient mental resources to allow them to notice target word forms. But, in Study 2, the increased task demands, induced by local-level task manipulation and the time limit, might have inhibited the participants from attending to the forms of glossed words. The so-called *involvement load* (Hulstijn & Laufer, 2001; Laufer & Hulstijn, 2001) could additionally be responsible for the negative effects of glossing on form recognition. That is, as the meanings were readily accessible from the glosses provided, participants could have chosen not to make deliberate attempts to work out the meanings of unknown words through inspection of word forms. In the case of Study 1, the unlimited time and relatively smaller cognitive demands might have allowed the

participants to allocate some mental resources to process pseudo-word forms, which was less practicable in Study 2.

In addition, glossing did not facilitate development in the knowledge of target unaccusative verbs in Study 2. The increased level of task complexity, combined with the time limit, could have overridden the influence of glossing on learning target verbs. Another possible scenario is that learners might not have been motivated to look at the glosses unless the target verbs interfered with reading comprehension or task completion. It should be noted that the participants had some prior knowledge of the meanings of target verbs. What they lacked was knowledge about unaccusative usage of the verbs (i.e., passive meaning in active voice, as in *The door opened*), not prototypical meanings (e.g., *I opened the door*). Thus, as long as the disparity between active voice and passive meaning of the target verbs did not interrupt their text understanding and/or task completion to a critical extent, the participants might have had little incentive to take advantage of the glosses. In the case of pseudo-words, however, the participants had absolutely no prior knowledge of these, which could have led to a greater need to refer to the glosses to obtain the meanings of those words.

4. Nature of the knowledge acquired

With respect to the nature of the knowledge acquired about the target grammatical features, it was found that, although reaction times in the GJT significantly decreased from a pretest to immediate and delayed posttests, neither task complexity nor glossing made a significant contribution to this trend. Also, the reaction time data revealed that it took significantly longer for the participants to react to ungrammatical sentences compared to grammatical ones. Again, this finding seemed fairly predictable, given that readers typically take more time when processing lexically and/or grammatically violated sentences, trying to resolve anomalies before ultimately marking them as

unacceptable or ungrammatical (Bley-Vroman & Masterson, 1989; Jiang, 2011; Juffs, 2001; L. White & Juffs, 1998).

Next, the analysis of confidence intervals and source attributions showed that, under the + complex condition, the participants were, overall, confident in their correct responses when they read unglossed texts, but less confident when they read glossed texts. This finding was diametrically opposite to that of Study 1. More specifically, under the + complex condition in Study 1, the participants were more confident in their responses to the GJTs when they read glossed texts, but not sure about their judgments when they read unglossed texts. These contrasting findings may be explained by the differential levels of the cognitive demands imposed under the + complex conditions in Studies 1 and 2. In Study 1, participants were allowed to take as much time as they needed, which might have left them with spare attentional resources to be shared out for processing the glosses. In Study 2, however, local task manipulation and the time limit might have led the participants to engage only in peripheral processing of the glosses. Additionally, when reading unglossed texts under the + complex condition, the participants might have had to figure out the passive meanings of target verbs in the active voice by themselves while ordering the sentences, and hence could be more confident when they were responding to GJT items.

When it comes to the nature of the knowledge acquired of pseudo-words, reaction time data revealed that it took significantly longer for the participants to respond to meaning than form recognition. Additionally, it was found that the sensitivity index *d'* was, overall, higher than zero for form recognition, while this was not the case for meaning recognition. In other words, participants were in general more confident about their correct answers for form recognition items, but doubtful even when they were correct for meaning recognition. The findings from reaction time data, together with the confidence intervals, seemed logical, given that processing word meanings requires not

only identifying visual patterns of perceived letters, which suffices in the case of processing word forms, but also producing phonemic recoding of registered sequences of letters and finding their semantic qualities from a lexical inventory. Thus, the acquisition of receptive knowledge of word forms must have required simpler processing than that of word meaning recognition, which accounts for the faster recognition and stronger confidence for form recognition in comparison to meaning recognition.

5. WMC as a mediator of the effects of task complexity and glossing

Last but not least, working memory capacity was analysed as a potential moderator of the effects of task complexity and glossing. First, when provided with glossed texts under the – complex condition, the participants with higher operation span scores were more likely to answer the reading comprehension items correctly than those with lower operation span scores. Thus, the results seemed to indicate that complex working memory plays an important part in storing a gloss and processing it for text understanding. When the glossed texts were provided in the + complex condition, complex working memory might have been used up by comparing and determining the correct order between sentences, and hence could not have moderating effects on glossing.

Some unexpected inverse relationships were also found. When reading Text 1 in the glossed version, the participants with higher digit span scores performed worse than those with lower digit span scores. Also, when reading Text 1 in the glossed version under the + complex condition, the participants with higher backward digit span scores performed worse than those who achieved lower scores. Recall that digit span scores and backward digit span scores shared significant correlation, $r_s(86) = .44$, $p < .01$, suggesting a partially overlapping function, probably the storage of information in short-term memory. That said, storage capacity could have been consumed by retaining

glosses as well as multiple propositions, which only had a negative influence on reading comprehension scores. However, it is still open to question why the participants with higher digit span scores or backward digit span scores were less able to manage additional information (i.e., glosses and more pieces of meaning units) than those with lower scores. Plus, the differential effects of working memory on reading comprehension scores for Texts 1 and 2 are also difficult to understand.

The results additionally revealed that, when assigned in the + complex and glossed condition, the participants with higher digit span scores obtained smaller GJT gain scores for target verbs in an immediate posttest than those with lower digit span scores. These unexpected results might be accounted for by a similar logic to that for the negative correlations between forward and backward digit span scores with reading comprehension scores. Namely, when the participants had to process glosses, while at the same time dealing with multiple propositions, storage capacity could have been depleted and hence exerted only a deteriorating effect on learning target verbs. Yet, again, why those with lower digit span scores could learn more about target verbs than those with higher digit span scores in the [+ complex, + glossing] condition remains unexplained.

Lastly, when performing the – complex tasks, the participants with higher operation span scores achieved smaller delayed meaning recognition scores than those with lower operation span scores. Perhaps the participants with higher operation span scores utilized their complex working memory more on the text-ordering task, rather than paying much attention to the target words. However, again, it remains inexplicable why those with lower operation span scores could gain higher vocabulary meaning recognition scores while performing the text-ordering task in the – complex condition. In sum, it seems clear that no conclusion can be drawn from the puzzling results of this study and that the inverse relationships warrant further empirical investigation in order

to determine whether and how working memory moderates the effects of task complexity and glossing on L2 reading comprehension and L2 learning.

VI. Unanswered Questions

Study 2 revealed several unresolved issues, particularly pertaining to the actual cognitive processes occurring in learners' minds when performing tasks that entail different features. First, while task complexity was shown to affect the unannounced retrospective and subjective time estimation in Study 1, as well as the perceived mental effort in Study 2, a more robust methodology, preferably susceptible to learners' online processes, seemed desirable in order to validate the construct of task complexity. In addition, findings from previous studies have demonstrated that task effects on L2 reading might be observable in online reading processes rather than in comprehension outcomes. Thus, proper process measures, such as verbal protocols and/or physiological techniques, were considered highly useful in delving into this issue. Also, it appeared worthwhile to explore whether task conditions had any influence on noticing target features while performing tasks. An ideal methodology for resolving these issues could be eye-tracking technology in combination with verbal protocols, which has become increasingly popular among SLA researchers as an accurate, robust and unobtrusive research methodology for inspecting online cognitive processes.

CHAPTER 5

STUDY 3

The present chapter reports an eye-tracking study, combined with stimulated recall protocols, which was conducted to address the remaining questions that emerged from Studies 1 and 2. In particular, the aim of this study was to explore whether task complexity manipulations might have an observable impact on learners' reading processes, which was not attested to in reading comprehension scores. Also, it was hoped that examining reading processes across different task complexity conditions would offer an additional way to assess the validity of the task complexity manipulation. Finally, the study sought to investigate how learners' processing of glosses might be affected by differential task demands. As such, the following research questions were addressed:

RQ (1) To what extent do the cognitive demands of second language reading tasks affect reading processes, as reflected in participants' eye-movements and stimulated recall comments?

RQ (2) To what extent do the cognitive demands of second language reading tasks affect the noticing of glossed linguistic constructions, as reflected in participants' eye-movements and stimulated recall comments?

This chapter is organized as follows. First, a detailed description of the research methodology is presented, including the research design, the participants, the research instruments, the procedures of the study and the analyses conducted on the data. The methodology section is followed by presenting the results from mixed-effects regression analyses on eye-movement measures and qualitative analysis of stimulated recall comments. The chapter concludes with a brief summary of the study and an interim discussion of the findings and limitations.

I. Methodology

1. Design

Thirty-eight L2 users of English participated in the study. They all completed two versions of an experimental reading task, a version that required the reading of Text 1 and a version that involved the reading of Text 2. The target constructions (see Tables 41 and 42 in Chapter 4) were glossed for both Text 1 and Text 2. Following a 2x2 repeated-measures design, participants were randomly assigned to four groups, depending on whether they were exposed to the two texts under a – complex and/or + complex task condition. Text order was also counterbalanced, resulting in four additional groups. Thus, the final experimental design comprised eight groups in total (A-H), as shown in Table 78.

Table 78. Experimental conditions for Study 3

Group	**Task 1** Text 1	**Task 2** Text 2	**Group**	**Task 1** Text 2	**Task 2** Text 1
A (*n*=5)	– C	– C	C (*n*=4)	+ C	+ C
B (*n*=5)	+ C	+ C	D (*n*=4)	– C	– C
E (*n*=5)	– C	+ C	G (*n*=5)	+ C	– C
F (*n*=5)	+ C	– C	H (*n*=5)	– C	+ C

As illustrated in Figure 17, all participants first completed a background questionnaire and completed a pretest and an English proficiency test. While they were carrying out the reading tasks, their eye-movements were recorded using an in-built eye-tracker. Each task was followed by a post-reading questionnaire. Immediately after performing the two reading tasks, eleven participants were asked to participate in a stimulated recall protocol. These students were randomly selected from among the groups that completed both the + and – complex task versions (groups: E, F, G, H).

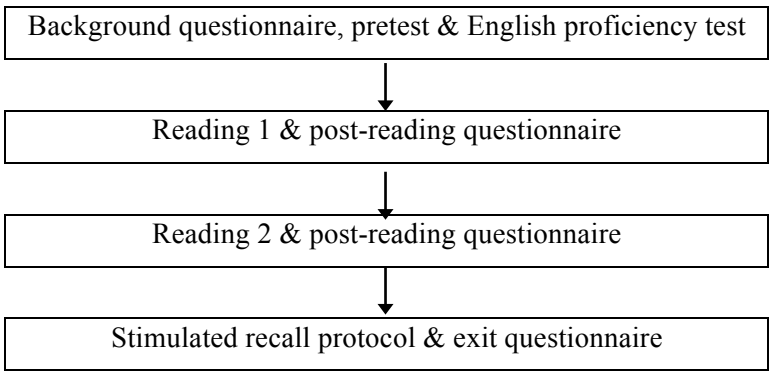

Figure 17. Procedure for Study 3

2. Participants

The 38 participating students were all L1 speakers of Korean. They were enrolled at a university in London, studying towards undergraduate (21%) or postgraduate degrees (79%). The ideal group of participants would have been Korean undergraduate students studying in Korea, as in Studies 1 and 2. However, as an eye-tracker was only accessible in the UK, this group of participants was chosen for the present study. Six students were male and 32 were female. Their ages ranged between 21 and 40 years old (Mean: 27.84, *SD* = 4.52). The average length of stay in an English-speaking country was 9.92 months (*SD* = 3.84). To ensure the homogeneity of the participants, their English proficiency level was measured with an adapted version of the *Use of English* section of a commercially available practice test: *Cambridge Proficiency: English* (CPE), developed and provided by the *University of Cambridge ESOL examinations*. The CPE test version used in the present study was the same as that used in Study 2. Cronbach's alpha for the CPE scores was .82

3. Reading tasks and target constructions

The same texts as in Studies 1 and 2 were used in this study (Text 1: Petroleum Resources; Text 2: The Cambrian Explosion). The + and – complex tasks were also taken from Study 2. That is, the participants were asked to determine the correct order of two subparts of each paragraph when reading the – complex version, but three to four

subparts of each paragraph when reading the + complex version. The same unaccusative verbs from the two passages and pseudo-word items were included as target constructions. Yet, unlike in Studies 1 and 2, the participants had to complete what was presented on a monitor (a paragraph followed by an ordering task and reading comprehension questions) before moving on to the next paragraph. In other words, participants in Study 3 were not allowed to move across paragraphs.

4. Task layout

Although care was taken to retain the task format used in Study 2, some modifications were made to accommodate the eye-tracking methodology, as presented in Figure 18. First, the reading tasks were reconstructed using 11-point, double-spaced Courier font. This font was chosen as each consonant and vowel has the same width, and 11-point is the largest size that enabled texts, glosses, and reading comprehension questions to appear together on the same screen. Line spacing was also increased from single to double for the text part in order to capture participants' eye-movements more accurately. In order to compensate for the relatively small font size (11-point) and thereby capture eye-movements more accurately, the reading tasks were projected on a 22-inch Dell monitor. This way, the font size could be increased upto 15-point (21px). Unlike in Study 2 where the participants were asked to write down the correct order of sub-parts of the text, here the participants were instructed to click the letters used to label the paragraphs (e.g. [A], [B], [C] and [D]) in the correct order. To facilitate the recording of participants' text-ordering decisions, each text sub-part started on a new line. The participants were allowed 25 minutes for task completion. The maximum score for reading comprehension was 10 points (1 point for each of 8 multiple-choice items and 2 points for an item requiring summary completion) for each text.

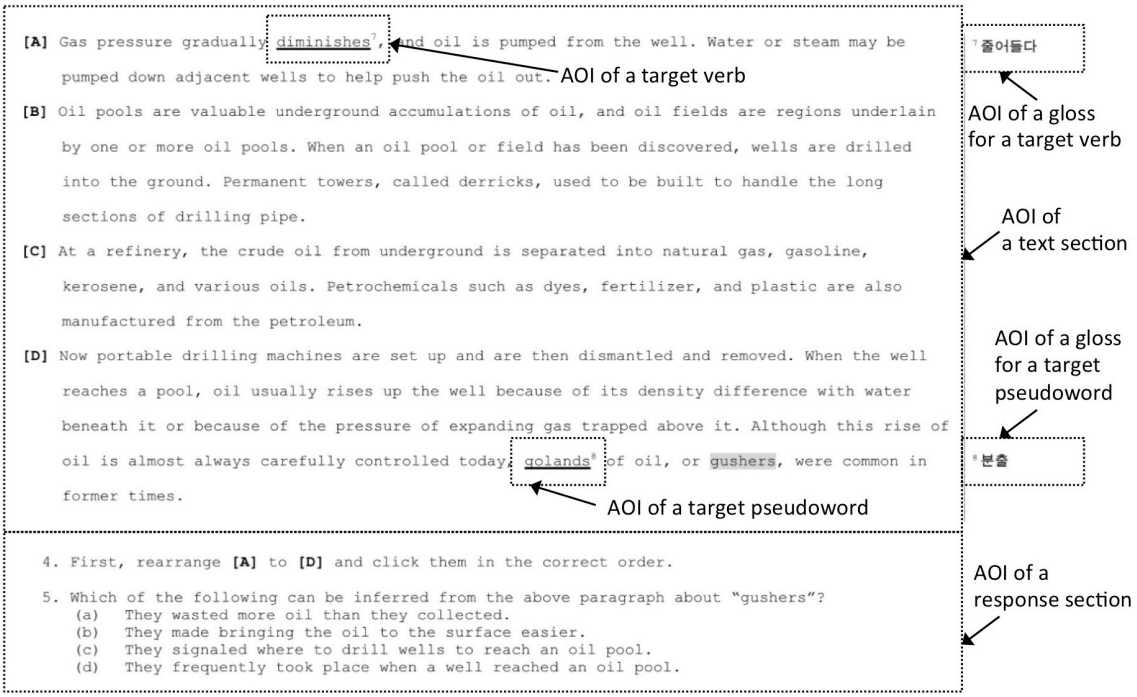

Figure 18. Sample task layout and areas of interest

5. Pretest

To measure the participants' prior knowledge of unaccusative verbs, the same grammaticality judgment test used in Study 2 was used at the outset of data collection. In this test, 16 sentences for novel unaccusative verbs were excluded, as learners' cognitive processes, not development in the knowledge of target constructions, were the focus of this study. As a result, the pretest contained 60 sentences in total. In other words, a grammatical judgment test included 15 grammatical and 15 ungrammatical sentences for the target unaccusative verbs and another 15 grammatical and 15 ungrammatical sentences to serve as distractors. The maximum score was 30, and the test took approximately 10 minutes to complete. Cronbach's alpha for the GJT scores was .62.

6. Stimulated recall

After completing both reading tasks, eleven students were asked to take part in stimulated recall sessions prompted by recordings of their eye movements. It was first explained to the participants that red lines and circles in the recordings indicated their

eye movements and fixation durations. They were also instructed to stop the recording at any time they wanted to verbalize what they were thinking while engaged in the original reading task. It was highlighted that they were supposed to report only what they were thinking at the moment of performing the tasks, not their interpretation of eye-movements at the time of verbalization. The researchers also interrupted the recordings and prompted the participants to describe their thoughts whenever unusual or interesting eye-movements were observed (longer fixation, regressive eye-movements or re-reading behavior) but these pauses were not commented on by the participants on their own (for the exact instructions, see Appendix F). The stimulated recall sessions were also video-recorded to capture the participants' spatial movements, especially when pointing at the computer monitor (e.g. *I started here, like this (pointing at screen), and it was very difficult*.)

7. Questionnaires

Participants were asked to complete a background questionnaire, a post-reading questionnaire and an exit questionnaire. The background questionnaire was designed to collect information about the participants' demographics and English learning experience. A post-reading questionnaire was administered after performing each reading task in order to assess the participants' perceived level of task difficulty and familiarity with the topics of the reading texts. Finally, an exit questionnaire asked the participants to provide retrospective comments about their experiences during task performance. All questionnaires were administered in Korean.

II. Procedure

Data were collected from two- to three-hour long sessions (see Figure 17). Participants first completed a background questionnaire, a pretest and an English proficiency test (CEP). Next, the eye-tracking system was calibrated, followed by a

brief session to help participants become familiar with doing computer-delivered reading tasks while sustaining an upright position for accurate detection of eye-movements. Then, the participants completed two reading tasks, each followed by a post-reading questionnaire. Eye-movements were captured with a mobile Tobii Pro X2-30 eye-tracking system with a temporal resolution of 30 Hz, which was mounted on the 22-inch Dell monitor. It should be noted that the sampling rate was considerably low for analysing reading processes with high precision, which is one of the limitations of Study 3. As such, the results of this study may need to be interpreted with caution, especially for the cases where the areas of interest were relatively small (i.e., noticing of the target constructions and their glosses). The participants were seated facing the eye-tracker approximately 60 cm from the center of the monitor, and their eye-movements were calibrated using a 9-point calibration grid. Immediately after finishing the reading tasks, the stimulated recall participants were further asked to recollect their reading processes in Korean, prompted by recordings of their eye-movements made during reading. Finally, each participant completed a questionnaire. Participants carried out the tasks individually in a quiet room at a university in London.

III. Analysis

This section presents analyses of eye-movement data and stimulated recalls. First, the procedure for extracting target eye-movement measures is introduced, followed by technical definitions for each of the measures and proposed hypotheses in relation to task complexity. Next, the statistical analyses employed to analyse eye-movement measures are described. Lastly, the steps taken to ensure the reliability of the transcribing and coding process for stimulated recalls are explained.

1. Eye-movement data

Eye-tracking data were analysed with Tobii Studio 3.0.9 (Tobii Technology, n.d.). For each page, areas of interest (AOI) were defined for (a) the text and (b) the text and response options combined (see Figure 18). Eye-movements captured on the text AOIs were used to extract indices associated with text reading processes, whereas AOIs for the text and response options combined were utilized as the basis for calculating measures of global processes during task performance. Then, drawing on Brunfaut and McCray's (2015) work, in total, ten indices of text and global processing were calculated based on eye fixation and saccade data obtained from Tobii Studio using R-script (McCray, personal communication, 9 August 2016). The global processing measures included more summative values, i.e., total frequencies and durations of eye-movements captured in both the texts and the response options combined, which were considered to reflect the participants' task performing processes as a whole. It was hypothesized that participants would spend more time on the task under the + complex condition, resulting in increased values for these eye-measures. By contrast, text reading measures included more specific indices, such as frequencies and durations of eye-fixations made on the text section only, forward and regressive eye-movements detected in the text section, and proportion of regressive movements, which were assumed to reveal how the participants read the texts. It was assumed that participants would engage in more careful and recursive reading when performing the + complex version, and this would be demonstrated in the increased values for forward saccades and regressive eye-movements, as well as total fixation counts and durations recorded in the text sections. Median length of forward saccades, however, was expected to decrease in the + complex condition, as participants' reading processes would be more frequently interrupted while processing the smaller text segments (see Table 79).

Next, to examine if task complexity affected participants' processing of target constructions and glosses, AOIs were defined for each target construction and gloss. While the target areas were identical in pixel size for pseudo-words and glosses, those for unaccusative verbs were inevitably dissimilar due to the different lengths of the target verbs. This did not confound the results of the present study, however, as the + and – complex versions included the same AOIs. Using these AOIs, eight additional eye-tracking measures were extracted using Tobii Studio. More specifically, number of fixations and sum of fixation durations were calculated for the target unaccusative verbs, the target pseudo-words, and their glosses, which were considered to reflect participants' noticing of those features. As presented in Table 80, it was hypothesized that when reading the texts more carefully under the + complex condition, the frequencies and durations of eye-fixations made on the target constructions and their related glosses would increase accordingly.

2. Statistical analyses

SPSS 22.0 (Statistical Package for the Social Sciences) for Mac was used to examine the reliability of the tests as well as to compute descriptive statistics for the data. Test reliability was determined using Cronbach's alpha. The level of significance for this study was set at an alpha level of $p < .05$. Mixed-effects models were constructed to examine whether there were any significant differences among the participants who performed the – versus the + complex versions, in terms of their English proficiency scores, pretest scores or perceptions of task-generated cognitive load. To do this, statistical program R, version 3.3.0, was used (R Development Core Team, 2016). And to examine if eye-movement indices differed significantly between the + and – complex versions, linear mixed-effects models were constructed using the *lmer* function provided by the *lme4* package (Bates, Maechler, & Bolker, 2012). All models included Complexity and Text as fixed effects, a random intercept for Subject

and a within-subject random slope for Complexity. Absolute *t*-values above 2.0 were set as the criterion for significance of the models (Gelman & Hill, 2007), and effect sizes were computed with the *r.squaredGLMM* function provided by the *MuMln* package (Barton, 2015). As suggested by Plonsky and Oswald (2014), R^2 above .06, .16 and .36 was considered as small, medium and large, respectively.

3. Stimulated recalls

The stimulated recall sessions were transcribed using the video-transcription software F5, version 2.2 (see Figure 19 for an illustration). The transcripts were uploaded to NVivo 10.0.3 software for qualitative analysis. The researcher reviewed the transcripts and identified emergent categories in a bottom-up manner by annotating the data. While there was no a priori coding scheme, the stimulated recalls were coded with the research questions (i.e., the effects of task complexity on the reading process and noticing) in mind. As data-driven annotations accumulated, they evolved into four major categories: participants' affective states, use of comprehension strategies, text-ordering task performance, and noticing of the target constructions or their glosses. The comments related to the participants' affective states subsumed comments related to high task demands, low task demands, and ability to concentrate on the task. Comments related to text comprehension included various reading strategies, such as re-reading, careful reading, and skimming. Annotations related to text-ordering task were further divided into word-level cues and discourse-level cues. Word-level cues involved participants' comments about relying on lexical cues, such as keywords, pronouns or articles, whereas discourse-level cues incorporated participants' recalls on how they analysed and compared sentence orders. The noticing category was further classified into annotations related to the target unaccusative verbs and those relevant to the target pseudo-words. Lastly, for each of the annotations, it was marked whether the comment

was produced under the + or – complex condition. The resulting coding scheme can be found in Table 92.

After coding all the transcripts, a randomly selected subset of the video-recordings (13.6%) was watched and coded by a second coder, an expert in Applied Linguistics, in order to verify the reliability of the coding. Agreement between the researcher and the second coder was 90 per cent with a kappa of .71 (*SE* = 1.02, 95% CI [- .98, 3.06]), which was acceptable. Next, comments were further categorized depending on whether they concerned the – or + complex condition, and frequency counts were calculated for each code under each condition.

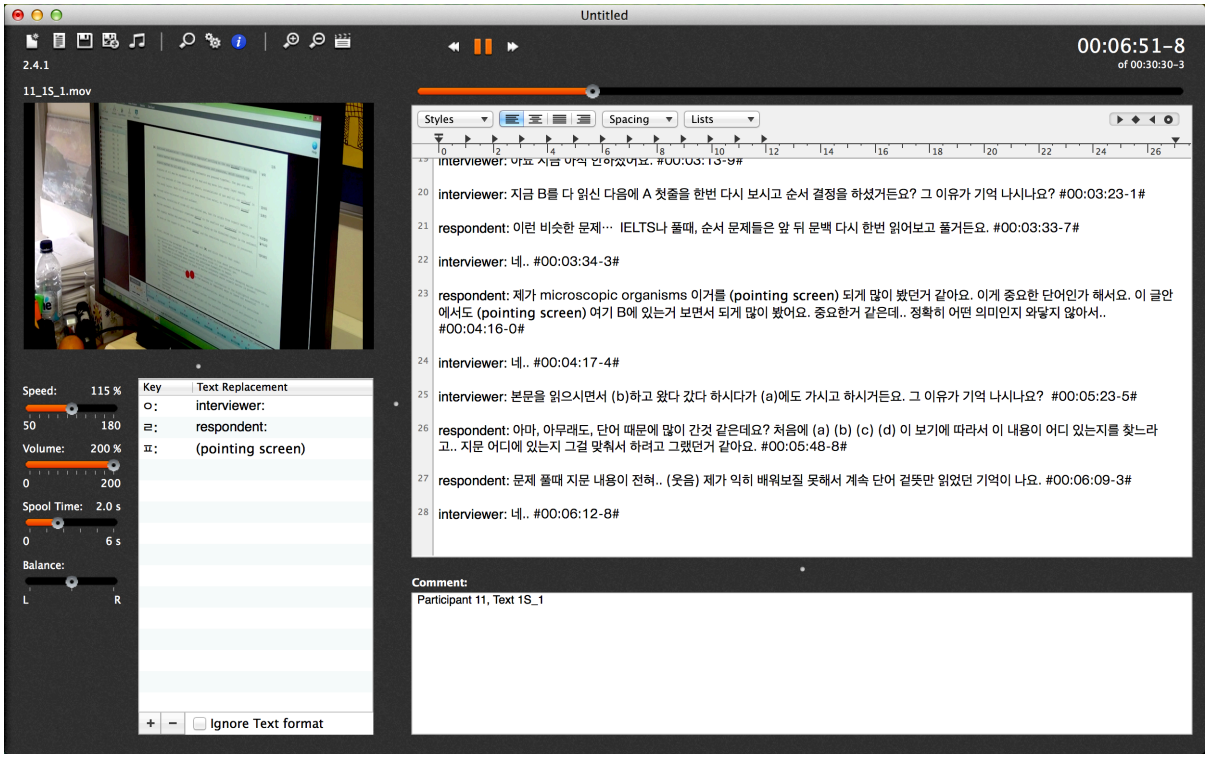

Figure 19. F5 video transcription

Table 79. Eye-movement measures and hypotheses for reading processes (adapted from Brunfaut & McCray, 2015)

Focus	Measure	Definition	Hypothesis
Global processing	Number of fixations	The sum of the number of fixations on texts and responses.	The number of fixations on tasks will be greater in the + complex versions, as more careful reading will be required.
	Sum of fixation durations	The sum of all fixation durations on texts and responses in seconds.	The sum of fixation durations on tasks will be longer in the + complex versions, as it will take longer to determine the order of text sub-parts.
Text reading	Number of fixations	The number of fixations on texts.	The number of fixations on texts will be greater in the + complex versions, as determining the correct order of text sub-parts will require more intensive textual processing.
	Sum of fixation durations	The sum of all fixation durations on texts.	The sum of fixation durations on texts will be longer in the + complex versions, as more careful reading will be required.
	Median fixation duration	Median fixation duration on texts, expressed in milliseconds.	The median fixation duration will be longer in the + complex versions, as more attentive textual processing will be required.
	Number of forward saccades	The number of forward saccades (eye-movements from point x to point y where point y lies to the left of point x).	The number of forward saccades will be greater in the + complex versions, as the texts will be processed more thoroughly.
	Median length of forward saccades	Median length, as expressed in pixels, of all forward saccades.	The median length of forward saccades will be shorter in the + complex versions, as textual processing will be more frequently interrupted due to increased cognitive load.
	Number of regressions	The number of regressions (eye-movements from point x to point y where point y lies to the right of point x)	The number of regressions will be greater in the + complex versions, as more repetitive and recursive textual processing will be necessary.
	Median length of regressions	Median length, as expressed in pixels, of all regressions	The median length of regressions will be greater in the + complex versions, as the complex task will entail more thorough processing of the text in order to confirm inter-sentential relations.
	Proportion of regressive movements	The number of regressions divided by the sum of the number of both forward saccades and regressions	The proportion of regressive movements will be greater in the + complex versions, as repetitive and recursive reading will be required to a greater extent.

Table 80. Eye-movement measures and hypotheses for noticing

Focus	Measure	Definition	Hypothesis
Verbs	Number of fixations	The number of all fixations on target verbs.	The number of fixations on target verbs will be greater in the + complex versions, as more intensive processing of the texts will result in repeated processing of verbs.
	Sum of fixation durations	The sum of all fixation durations on target verbs.	The sum of fixation durations on target verbs will be longer in the + complex versions, as more careful reading of the texts will involve increased exposure to the verbs.
Glosses for verbs	Number of fixations	The number of all fixations on glosses for target verbs.	The number of fixations on the glosses for target verbs will be greater in the + complex versions, as + complex version will require accurate understanding of each sentence.
	Sum of fixation durations	The sum of all fixation durations on glosses for target verbs.	The sum of fixation durations on glosses for the target verbs will be longer in the + complex versions, as the glosses will more likely be processed in the course of more attentive processing of texts.
Pseudo-words	Number of fixations	The number of all fixations on pseudo-words.	The number of fixations on pseudo-words will be greater in the + complex versions, as they will be processed more frequently in the course of performing text-ordering tasks.
	Sum of fixation durations	The sum of all fixation durations on pseudo-words.	The sum of fixation durations on pseudo-words will be longer in the + complex versions, as more careful reading of the texts will result in longer eye-gazes on words.
Glosses for pseudo-words	Number of fixations	The number of all fixations on glosses for pseudo-words.	The number of fixations on glosses for pseudo-words will be greater in the + complex versions, as the reordering of sentences may increase the need to check the meanings of words.
	Sum of fixation durations	The sum of all fixation durations on glosses for pseudo-words.	The sum of fixation durations on glosses for pseudo-words will be longer in the + complex versions, as more thorough processing of the texts will entail more in-depth processing of the word meanings provided in the glosses.

IV. Results

1. Preliminary analysis

Prior to answering the research questions, some preliminary steps were taken to ensure the reliability and validity of the results. The following methodological concerns were taken into consideration: prior knowledge of target constructions, the impact of topic familiarity on reading comprehension scores, and perceived level of task complexity.

1.1. Equivalence by task complexity and text conditions

To check the equivalence of the English proficiency level of the participants, multi-level mixed-effects models were constructed. CPE scores were the dependent variable, and null models included Subject and Item as random effects, and Complexity and Text as within-subject random slopes for Subject. To this null model, Complexity and Text were entered one by one as a fixed effect. The results showed that the inclusion of neither Complexity nor Text made a significant difference to the null model, Complexity: $\chi^2(1) = .01$, $p = .93$, $R^2 < .01$, Text: $\chi^2(1) = .01$, $p = .99$, $R^2 < .01$, Complexity*Text: $\chi^2(1) < .01$, $p = .99$, $R^2 < .01$. In other words, the proficiency level of the participants did not change significantly as a function of text and task complexity condition (for descriptive statistics, see Table 81).

Table 81. Descriptive statistics for proficiency test

	– Complex (n = 19)		+ Complex (n = 19)	
	Mean	*SD*	Mean	*SD*
Text 1	11.40	5.05	9.22	4.81
Text 2	9.53	4.65	11.00	5.22
Total	10.40	5.01	10.10	5.11

Note. Maximum score = 30. Each score was calculated twice due to the experimental design.

Next, to test whether the participants started out at a developmentally parallel stage, another set of likelihood ratio tests were conducted on the pretest GJT scores (for descriptive statistics, see Table 82). The null model included Subject and Item as random effects, and Complexity and Text as within-subject random slopes for Subject.

Increased models additionally contained Complexity and Text as fixed effects. The results indicated that Complexity and Text did not improve the null models to a significant degree, Complexity: $\chi^2(1) = .01$, $p = .93$, $R^2 < .01$, Text: $\chi^2(1) = .01$, $p = .99$, $R^2 < .01$, Complexity*Text: $\chi^2(1) < .01$, $p = .99$, $R^2 < .01$. In sum, the results showed that, at the time of the pretest, there were no significant differences among the participants in terms of their ability to judge the grammaticality of English unaccusative sentences across text and task complexity allocation.

Table 82. Descriptive statistics for pretest scores

	– Complex (n = 19)		+ Complex (n = 19)	
	Mean	*SD*	Mean	*SD*
Text 1	15.74	4.57	14.95	3.52
Text 2	15.39	4.36	15.53	4.02
Total	15.50	4.38	15.43	3.87

Note. Maximum score = 30. Each score was calculated twice due to the experimental design.

1.2. Effects of topic familiarity

To confirm that the participants' topic knowledge did not affect their reading, the participants' familiarity with the two topics was measured using post-reading questionnaire items (e.g. Item 6: *I thought this topic of the reading was familiar*, Item 13: *I had some background knowledge about the reading topic*). The descriptive statistics are presented in Table 83. The responses to the two items correlated significantly with each other, Text 1: $r(38) = .80$, $p < .01$, Text 2: $r(38) = .87$, $p < .01$, suggesting that the items assessed overlapping constructs. A series of likelihood ratio tests found that adding Familiarity and its interactions with the fixed effects did not significantly improve the null model, Familiarity: $\chi^2(1) = 1.56$, $p = .21$, $R^2 < .01$, Complexity * Familiarity: $\chi^2(1) = .89$, $p = .34$, $R^2 < .01$, Text*Familiarity: $\chi^2(1) = 1.47$, $p = .22$, $R^2 < .01$, Familiarity*Complexity*Text: $\chi^2(4) = 5.37$, $p = .25$, $R^2 = .01$. In short, topic familiarity did not affect participants' reading comprehension scores.

Table 83. Descriptive statistics for topic familiarity by item

		Text 1		Text 2	
Item	*N*	*Mean*	*SD*	*Mean*	*SD*
# 6	38	2.34	1.48	2.08	1.19
#13	38	2.34	1.30	2.03	1.22
Total	38	2.35	1.39	2.05	1.19

Note. Maximum value: 7.

1.3. Validation of task complexity manipulation

To infer the effects of task complexity on the level of cognitive load imposed on the participants, three questionnaire items were included in post-task questionnaires (Item 1: *I thought this task was difficult*, Item 7: *I invested a large amount of mental effort to complete this task*, Item 14: *I thought this task was demanding*). Cronbach's alpha for the items was .83 for Text 1 and .75 for Text 2. The descriptive statistics presented in Table 84 show that, overall, the participants perceived the + complex task as slightly more demanding. The results of likelihood ratio tests comparing null models with random effects (i.e., Subject and Item) only and increased models additionally containing either Text or Complexity revealed that there was no significant difference across the task conditions, Text : $\chi^2(1) = .07$, $p = .79$, $R^2 < .01$; Complexity: $\chi^2(1) = .73$; $p = .39$, $R^2 = .02$. In short, participants' perceived level of mental effort appeared comparable regardless of task manipulation.

Table 84. Descriptive statistics for perceived task difficulty by item

			Reported mental effort					
			Text 1			Text 2		
Item	Condition	*N*	*Mean*	*SD*	*SE*	*Mean*	*SD*	*SE*
# 1	– Complex	19	5.05	1.27	.29	5.32	.95	.22
	+ Complex	19	5.68	1.06	.24	5.63	1.12	.26
# 7	– Complex	19	5.05	1.22	.28	4.53	1.50	.35
	+ Complex	19	4.53	1.74	.40	4.89	1.41	.32
# 14	– Complex	19	4.79	1.47	.34	4.63	1.30	.30
	+ Complex	19	5.05	1.43	33	5.11	1.05	.24
Total	– Complex	19	14.89	4.38	1.01	14.48	4.25	.97
	+ Complex	19	15.26	4.47	1.03	15.63	5.11	1.17

Note. Maximum value for each item = 7.

2. Eye-movement data

This section reports the results for eye-movement measures, which were obtained to gain insights into the cognitive processes in which the participants engaged. First, results are presented for the extent to which task complexity affected the participants' reading processes, followed by reports on how the noticing of target constructions differed between the – versus + complex conditions.

2.1. Task complexity and eye-movements related to reading processes

Table 85 presents descriptive statistics for eye-movement measures related to reading processes. A glimpse at the table reveals larger values in the + complex condition for several eye-movement measures, namely, number of fixations for texts and responses combined, number of fixations for texts only, sum of fixation durations for texts only, number of forward saccades and number of regressions. Also, under the + complex condition, sum of fixation durations capture on both texts and responses seemed longer for Text 2 than Text 1.

Table 85. Descriptive statistics for eye-movement measures of reading processes

	Global processing		Text-reading							
	Number of fixations	Sum of fixation durations	Number of fixations	Sum of fixation durations	Median fixation duration (ms)	Number of forward saccades	Median length of forward saccades (px)	Number of reg-ressions	Median length of reg-ressions (px)	Propor-tion of regressive movements
- Complex										
Text 1										
Mean	2836.74	739.13	1570.84	426.50	221.11	1024.53	96.42	403.53	-164.50	0.28
SD	564.66	176.81	387.51	108.56	28.82	265.82	9.47	127.02	42.14	0.04
95% CI Low	2589.65	659.50	1397.66	375.40	208.26	912.34	92.44	349.01	-182.22	0.26
95% CI Up	3102.04	817.25	1753.23	477.32	234.36	1149.32	100.82	463.99	-146.59	0.30
Text 2										
Mean	2893.58	775.73	1457.68	405.42	225.21	920.11	95.55	400.58	-163.37	0.30
SD	602.12	202.67	317.13	115.38	34.12	192.94	10.37	129.73	47.31	0.05
95% CI Low	2641.97	695.76	1317.84	359.33	210.76	829.82	91.14	345.35	-183.58	0.28
95% CI Up	3181.36	870.48	1603.83	463.83	240.55	1000.88	100.00	465.37	-143.11	0.32
+ Complex										
Text 1										
Mean	3120.95	597.09	1894.21	500.29	216.53	1152.74	97.71	501.74	-154.58	0.30
SD	588.73	151.41	379.76	121.05	40.40	271.31	14.42	154.87	49.24	0.05
95% CI Low	2874.44	532.49	1723.48	442.90	197.58	1036.15	91.77	435.69	-176.66	0.28
95% CI Up	3376.99	661.88	2049.78	549.46	232.89	1276.35	104.29	569.47	-132.88	0.32
Text 2										
Mean	3487.79	861.94	2112.37	542.36	211.32	1317.32	96.05	536.26	-167.39	0.29
SD	694.27	201.21	529.07	146.73	36.09	332.42	12.10	182.00	49.60	0.04
95% CI Low	3169.48	768.09	1878.34	469.44	195.26	1155.04	90.54	463.08	-191.52	0.27
95% CI Up	3788.66	942.47	2355.19	602.22	227.26	1457.02	101.85	622.14	-147.18	0.31

Next, a series of likelihood ratio tests were administered to test whether the participants' reading processes, as reflected in eye-movement measures, were influenced by different levels of task complexity. The null models included Subject as a random effect. To this null model, Complexity and Text were added one by one as an additional fixed effect to see if its inclusion improved the fit of the null models to a significant extent. As summarized in Table 86, Text improved the null model significantly for the sum of fixation durations on the text and response parts combined. Also, Complexity emerged as a significant predictor of the following measures: number of fixations for texts and responses combined, number of fixations and sum of fixation durations for texts only, number of forward saccades and number of regressions. For these cases, post hoc mixed-effects models were constructed to further examine the influence of Complexity and Text on participants' eye-movements. Lastly, a significant interaction effect was also found for proportion of regressions ($\chi^2(1) = 4.17$, $p = .04$, $R^2 = .03$). No significant effects were observed for the rest of the measures.

Table 86. Significant results from likelihood ratio tests for eye-movement measures of reading processes

Fixed-effect	Measure	χ^2	*df*	*p*	R^2
Text	Number of fixations_Text & Response	3.69	1	.05$^+$	.03
	Sum of fixation durations_Text & Response	23.63	1	< .01**	.14
	Number of fixations_Text	.35	1	.56	.00
	Sum of fixation durations_Text	.18	1	.67	.00
	Number of forward saccades	.01	1	.90	.00
	Median length of forward saccades	.31	1	.58	.00
	Number of forward regressions	2.71	1	.10	.00
	Median length of regressions	.35	1	.55	.00
	Proportion of regressions	.17	1	.68	.00
Complexity	Number of fixations_Text & Response	13.39	1	< .01**	.12
	Sum of fixation durations_Text & Response	.15	1	.70	.03
	Number of fixations_Text	29.46	1	< .01**	.27
	Sum of fixation durations_Text	16.91	1	< .01**	.17
	Number of forward saccades	23.25	1	< .01**	.21
	Median length of forward saccades	3.18	1	.07$^+$	.01
	Number of regressions	18.26	1	< .01**	.16
	Median length of regressions	.80	1	.37	.00
	Proportion of regressions	.79	1	.37	.00

Note. Significance level: $^+p < .1$, $*p < .05$, $**p < .01$.

Table 87 presents the results for post hoc multi-level mixed-effects models of eye-gaze measures. The summaries of post hoc mixed-effects models confirmed the significant influence of Text and Complexity on the eye-movement measures. To be more specific, it took significantly longer for the participants to complete the tasks that included Text 2 in comparison to Text 1. When it comes to the text-sections only, it took significantly longer to complete the + complex tasks than the – complex ones, as manifested in the sum of fixation durations captures on the texts. The number of fixations was significantly greater in the + complex tasks for both task as a whole and text parts only. In addition, the participants made significantly more amounts of forward as well as regressive eye-movements in the + complex versions. In other words, under the + complex condition, they appeared to engage in more repetitive and recursive reading, as manifested in the increased numbers of forward saccades and regressions. Turning to the interactions, greater task complexity was found to increase the sum of fixation durations on both text and response sections in Text 2 but decreased the index in Text 1. Also, increased task complexity affected proportion of regressions positively in Text 1, but negatively in Text 2. The R^2 values for these relationships ranged from .12 to .22, indicating small to medium effect sizes. The only exception was a very small effect size (R^2 = .03) observed for the Interaction on the proportion of regressions.

Table 87. Summary of mixed-effects models for eye-movement measures of reading processes

	Fixed effects			Random effects	
				by subject	by text *or* by complexity:subject
	Estimate	*SE*	*t*	*SD*	*SD*
Intercept	743.47	28.00	26.55°	151.80	114.40
Text	150.72	26.68	5.65°	–	–
SumFixationDurationTextResponse ~ Text + *(1| Subject); R^2 = .14.*					
Intercept	3084.76	87.96	35.07°	449.10	.00
Complexity	463.57	119.99	3.86°	–	–
NumberFixationsTextResponse ~ Complexity + *(1| Subject) + (1|Complexity:Subject); R^2 = .12.*					
Intercept	1758.78	58.37	30.13°	294.10	.00
Complexity	496.96	81.37	6.11°	–	–
NumberFixationsText ~ Complexity + *(1| Subject) + (1|Complexity:Subject); R^2 = .17.*					
Intercept	468642	17110	27.39°	83718	.00
Complexity	109063	24851	4.34°	–	–
SumFixationDurationsText ~ Complexity + *(1| Subject) + (1|Complexity:Subject); R^2 = .14.*					
Intercept	1103.67	39.44	27.98°	205.00	.00
Complexity	279.29	52.09	5.36°	–	–
NumberForwardSaccadeText ~ Complexity + *(1| Subject) + (1|Complexity:Subject); R^2 = .21.*					
Intercept	460.53	21.52	21.41°	114.13	.00
Complexity	126.73	27.20	4.66°	–	–
NumberRegressionsText ~ Complexity + *(1| Subject) + (1|Complexity:Subject); R^2 = .16.*					
Intercept	743.47	27.29	27.24°	151.70	–
Complexity*Text	150.72	62.44	3.76°	–	–
*SumFixationDurationTextResponse ~ Complexity*Text + (1| Subject); R^2 = .22.*					
Intercept	.29	.01	41.95°	.04	–
Complexity*Text	-.03	.01	-2.01°	–	–
*ProportionRegressions ~ Complexity*Text + (1| Subject); R^2 = .03.*					

Note. Significance: °| *t* | > 2.0.

2.2. Task complexity and eye-movements related to noticing

The eye-movement measures associated with the noticing of target constructions as well as glosses were also analysed to see if task complexity had any influence on the indices. Table 88 presents statistics for the eye-movement measures for noticing, by task conditions and by texts. According to the table, the number of fixations and the sum of fixation durations for the target verbs appeared to be greater under the + complex condition.

Table 88. Descriptive statistics for eye-movement measures of noticing

	Verb		Verb gloss		Pseudo-word		Pseudo-word gloss	
	Number of fixations	Sum of fixation durations	Number of fixations	Sum of fixation durations	Number of fixations	Sum of fixation durations	Number of fixations	Sum of fixation durations
- Complex								
Text 1								
Mean	30.68	7.94	3.42	0.89	18.58	4.97	2.26	0.49
SD	12.21	2.85	2.61	0.75	8.44	2.36	2.64	0.63
95% CI Low	25.58	6.66	2.32	0.56	15.11	4.02	1.11	0.23
95% CI Up	36.31	9.18	4.53	1.21	22.58	6.06	3.42	0.78
Text 2								
Mean	38.05	11.45	3.37	0.87	20.26	5.25	2.79	0.62
SD	8.26	3.34	1.80	0.78	7.89	2.06	2.42	0.53
95% CI Low	34.58	10.04	2.58	0.58	17.00	4.41	1.74	0.39
95% CI Up	41.58	12.96	4.16	1.28	24.05	6.28	3.89	0.86
+ Complex								
Text 1								
Mean	43.95	11.79	3.63	0.70	20.89	5.40	2.63	0.51
SD	9.66	3.56	3.08	0.64	9.39	2.63	2.24	0.53
95% CI Low	39.90	10.24	2.37	0.45	17.11	4.31	1.69	0.29
95% CI Up	47.84	13.34	5.00	0.98	25.11	6.55	3.68	0.77
Text 2								
Mean	57.26	15.51	3.58	0.82	19.11	4.89	2.32	0.57
SD	21.82	6.93	3.78	1.16	8.11	2.56	3.23	0.95
95% CI Low	47.90	12.61	2.16	0.38	15.84	3.95	1.05	0.21
95% CI Up	67.47	18.54	5.31	1.34	22.74	6.01	3.84	1.02

Next, a set of likelihood ratio tests was conducted to examine whether Complexity and Text had a significant effect on these eye-movement indices. The null models included Subject as a random effect, and each of Complexity and Text were added to the null models to compare the model fit between the null and the increased models. As presented in Table 89, the results of likelihood ratio tests indicated that both Complexity and Text improved the model fit significantly for the number of fixations and sum of fixation durations captured on the target unaccusative verbs. The interaction, however, did not improve model fit (number of fixations: $\chi^2(1) = .97$, $p = .32$, $R^2 = .34$, sum fixation durations: $\chi^2(1) < .01$, $p = .93$, $R^2 = .28$).

Table 89. Significant results of likelihood ratio tests for eye-movement measures of noticing

Fixed-effect	Measure	χ^2	*df*	*p*	R^2
Text	Number of fixations_Verbs	7.88	1	< .01**	.09
	Sum of fixation durations_Verbs	10.89	1	< .01**	.12
	Number of fixations_Verbs_glosses	.01	1	.92	.00
	Sum of fixation durations_Verbs_glosses	.06	1	.80	.00
	Number of fixations_Words	.00	1	.98	.00
	Sum of fixation durations_Words	.05	1	.83	.00
	Number of fixations_Words_glosses	.05	1	.83	.00
	Sum of fixation durations_Words_glosses	.65	1	.42	.00
Complexity	Number of fixations_Verbs	20.44	1	< .01**	.23
	Sum of fixation durations_Verbs	12.50	1	< .01**	.15
	Number of fixations_Verbs_glosses	.36	1	.55	.00
	Sum of fixation durations_Verbs_glosses	.26	1	.61	.00
	Number of fixations_Words	.18	1	.67	.00
	Sum of fixation durations_Words	.00	1	.95	.00
	Number of fixations_Words_glosses	.02	1	.87	.00
	Sum of fixation durations_Words_glosses	.01	1	.91	.00

Note. Significance level: $^{+}p < .1$, $*p < .05$, $**p < .01$.

As Table 90 illustrates, the summaries of the multi-level mixed-effects models confirmed Complexity and Text as significant predictors of the number of fixations and the sum of fixation durations on the target unaccusative verbs. The effect sizes were medium, as reflected in $R^2 = .34$ for number of fixations and .28 for sum of fixation durations. In other words, the participants fixated significantly more often and longer on target verbs when performing the complex task versions compared to simple versions. In addition, the target verbs were more likely to be noticed in Text 2 than in Text 1.

Table 90. Summary of mixed-effects models for eye-movement measures of noticing

	Fixed effects			Random effects	
				by subject	by complexity:subject
	Estimate	*SE*	*t*	*SD*	*SD*
Intercept	42.49	1.77	24.01°	6.45	.00
Complexity	16.60	3.13	5.31°	–	–
Text	10.34	2.86	3.62°	–	–
Complexity*Text	6.07	6.32	.96	–	–
NumberFixationsVerbs ~ Complexity*Text + *(1 | Subject) + (1 |Complexity:Subject); R^2 = .34.*					
Intercept	11.68	.57	20.58°	1.99	1.15
Complexity	4.11	1.01	4.08°	–	–
Text	3.62	.90	4.04°	–	–
Complexity*Text	.18	1.97	.09	–	–
SumFixationDurationsVerbs ~ Complexity*Text + *(1 | Subject) + (1 |Complexity:Subject); R^2 = .28.*					

Note. Significance: °| *t* | > 2.0.

In sum, the results of mixed-effects modelling for eye-movement measures indicated that different levels of task complexity affected the participants' reading processes as well as their noticing of target unaccusative verbs. That is, when task complexity increased, the participants fixated more frequently and longer on the text, and they also seemed to engage in more intensive processing of it. In addition, the target verbs appeared to be processed to a significantly greater extent when participants performed the + complex versions.

3. Stimulated recall protocols

This section describes the coding scheme for the stimulated recalls and reports the results of frequency analysis. As illustrated in Table 91, eight meta-codes were identified: *High task demands, Low task demands, Ability to concentrate on task*, *Comprehension, Word-level cue*, *Discourse-level cue, Noticing of target verbs, and Noticing of target pseudo-words*. Each meta-code was further broken down into various sub-codes, and example comments are listed for each code in the table.

As presented in the table, more annotations were counted for the + complex (n = 374) than the – complex versions (n = 230), and this trend was seen for most of the codes. The participants produced greater numbers of comments related to feeling the given task difficult for the + complex version. Also, they more frequently expressed feeling unconfident about their task performance when recalling their thoughts for the + complex version. When commenting on their performance under the + complex condition, participants also mentioned more often that they encountered some comprehension difficulty and used various reading strategies, such as searching for hints, skimming a given text and reading texts carefully. On a word-level, the participants additionally reported that they utilized keywords, pronouns and transitional words with greater frequency under the + complex condition. On a discourse-level, they also tended to focus on the first sentence of each sub-part to a greater extent when

provided with the + complex version. They additionally mentioned that they struggled to put the text segments in order under the + complex condition. Lastly, the participants produced more recalls related to noticing target unaccusative verbs as well as related glosses when commenting on the + complex tasks.

Some inverse patterns were also found. For example, more annotations were marked in relation to rereading behaviour for the – complex condition. In addition, the participants more often reported that they focused on articles, first mentions of words and sentence connections when reading the – complex versions of texts. When it came to noticing, the participants generated more comments related to noticing glosses for pseudo-words when reading the – complex versions.

In sum, the stimulated recall protocols revealed that the participants engaged to some extent in different reading processes under the + and – complex conditions. The participants also more often reported experiencing difficulty and concentration problems when doing the more complex version, indicating that task manipulation was successful. Finally, when reading the + complex versions, the participants seemed more likely to process the target verbs and the related glosses.

Table 91. Code frequency for stimulated recalls (*n* = 604)

Meta-code/code	Complex (*n* = 374)	Simple (*n* = 230)	Example
High task demands	*57*	*13*	
Difficulty (High)	43	*9*	*It wasn't easy at all.*
Unconfident task completion	14	*4*	*I wasn't sure about my text ordering.*
Low task demands	*7*	*11*	
Difficulty (Low)	7	*8*	*It wasn't that difficult.*
Confident task completion	0	*3*	*I was thinking that I understood the content well.*
Ability to concentrate on task	*13*	*11*	
Concentration (Low)	11	*9*	*I could not concentrate well on the task in the beginning.*
Concentration (High)	2	*2*	*I could concentrate better on the task this time.*
Comprehension	*122*	*88*	
Overall comprehension	24	*25*	*I could not understand (A) when I first read it.*
Re-reading	20	*25*	*I tried to read this again.*
Careful reading	24	*15*	*I thought (B) came first, so I had to understand (B) perfectly before reading (A).*
Skimming	*22*	*12*	*I didn't read carefully, because I just wanted to see the overall structure.*
Searching for hints	26	*8*	*I was trying to find something that connects these text segments.*
Refer to previous passage	6	*3*	*I was thinking about the content of the previous passage.*
Word-level cues	*84*	*42*	
Keyword	40	*14*	*I thought "soft-bodied animal" was the keyword here.*
Signal word	18	*8*	*I assumed "finally" must indicate the last part of the text.*
Pronoun	14	*3*	*It wasn't the first, because it follows "these".*
Second mention	8	*5*	*I saw some repeated words. Repeated words were useful when deciding on order.*
First mention	2	*6*	*This was the first time "drilling" was mentioned.*
Article	2	*6*	*For instance, "a" became "the".*
Discourse-level cues	*78*	*56*	
Logical flow	24	*23*	*(B) gave a general statement, while (A) gave a concrete example.*
Wrestle to order segments	33	*11*	*I was debating about the order between these two segments.*
First sentence	17	*11*	*I thought focusing on the first sentence would be enough to decide on the order.*
Sentence connection	3	*9*	*I was checking if (A)-final and (B)-front, or (B)-final and (A)-front were connected.*
Final sentence	1	*2*	*If the sentences were connected, I thought there must be a clue in the final sentence.*
Noticing – Target unaccusative verbs	*9*	*4*	
Noticing glosses	6	*4*	*I could notice the glosses naturally, as they were in Korean.*
Noticing target verbs	3	*0*	*I thought "diminish" might be an important word here.*
Noticing – Target pseudo-words	*4*	*5*	
Noticing glosses	2	*5*	*The gloss for "golands" helped me to learn that it had a different meaning from "gusher".*
Noticing target words	2	*0*	*It was my first time seeing this word.*

V. Interim Discussion

The present study was designed to delve into some of the unresolved issues found in Study 2, namely, learners' cognitive processes while performing reading tasks with different levels of cognitive demand, which could not be attested in off-line learning scores or reading comprehension scores. To be more specific, it was expected that exploring learners' internal processes could provide empirical evidence regarding the validity of task complexity manipulation as well as the noticing of glossed target constructions. Eye-tracking technology, triangulated with stimulated recall protocols, was employed to investigate learners' reading processes during task performance. In this study, Korean speakers learning English performed + and/or – complex versions of reading tasks which required answering reading comprehension questions. The target constructions (i.e. English unaccusative verbs and pseudo-words) were glossed by means of providing Korean definitions in the margins of the texts. During task performance, the participants' eye-movements were recorded with an in-built eye-tracker, and eleven participants further participated in stimulated recall protocols while viewing their own eye-movements.

1. Task complexity and L2 reading processes

The eye-movement results suggest that the participants processed the texts more thoroughly under the + complex than the – complex condition, as reflected in the number of eye-gaze measures. The participants tended to fixate more on tasks when performing the + complex versions. In addition, they fixated more and for longer on the texts under the + complex condition, as manifested in the significantly larger number of fixations and longer fixation durations for the texts. The numbers of forward saccades and regressive eye-movements further indicate that the participants engaged in more attentive and recursive processing of texts. As discussed in Study 2, successful completion of the + complex tasks may have required closer inspection of the texts in

order to arrive at an accurate understanding of each sentence, as well as the logical relationships among them. Consequently, the participants might have had to read the texts more carefully and thoroughly when completing the + complex tasks, which was confirmed by the eye-movement data.

The analysis of stimulated recalls provided results compatible with eye-movements. On a global level, the participants reported that they perceived the + complex task as more demanding. In particular, they more often recalled wrestling to order the segments, and being unconfident about task completion. The significantly greater number of fixations during the task may represent the participants' deliberate endeavours to process the text for accurate understanding, which was crucial to order the text segments coherently under the + complex condition. The participants' comments also revealed that, under the + complex condition, they more frequently employed various reading strategies, such as skimming, careful reading and searching for hints. They also recalled more extensive use of lexical cues, including keywords, signal words, pronouns and words that were mentioned for a second time. That is to say, they appeared to process the texts more intensively using diverse metacognitive strategies under the + complex condition, which seems consistent with the longer fixation durations as well as the increased numbers of fixations, forward saccades and regressions captured in the texts.

It seems noteworthy, however, that for some of the eye-movement measures, no significant difference (median of fixation duration, median length of forward saccades, median length of regressions, and proportion of regressions) or an interaction effect (sum of fixation durations for text and response options combined and proportion of regressive movements) was found between the two task conditions, contrary to the hypotheses. A possible explanation why task complexity had no impact on the median measures may lie in that, although the two task versions led to a differential amount or

quantity of processing, they did not prompt qualitatively different reading processes. Medians of fixation, saccade or regression lengths or proportion of regressions might be likely to capture qualitative differences in reading processes. For example, longer saccade lengths are probably more associated with global rather than local reading, since global reading necessitates less detailed comprehension (Brunfaut & McCray, 2015). Also, a gap-fill task with a given set of words, for instance, would be more likely to increase proportion of regressions, as readers would probably revisit the list of words during task completion. The stimulated recall data support the account that the task complexity manipulation had primarily quantitative effects on reading processes: participants recalled using certain strategies with greater frequency under the + complex condition, but rarely mentioned the use of qualitatively different strategies.

When the issue turns to the interaction identified for sum of fixation durations for text and responses combined, participants fixates shorter overall during the + complex as compared to the – complex version for Text 1, while the pattern was in the expected direction for Text 2 with longer overall fixation duration in the + complex condition. A possible clue lies in the fact that participants achieved considerably lower mean scores on the text-ordering task on the + complex version of Text 1 (Mean =1.00) than Text 2 (Mean =1.95), suggesting that increasing task complexity resulted in proportionately greater demands for Text 1 than for Text 2. This might have left less attention available for answering the reading comprehension questions based on Text 1, which, in turn, might have led to shorter fixations on the Text 1 comprehension questions (but not the text itself). This account is consistent with the fact that sum of fixation durations for text only were, just as for Text 2, higher for the + complex than the – complex version of Text 1. Another possible explanation is that the + complex text-ordering task for Text 1, which appeared to be even more demanding than that for Text 2, encouraged

participants to engage in the text more thoroughly and repeatedly, resulting in quicker completion of the reading comprehension items.

2. Task complexity and noticing of glossed linguistic constructions

When it comes to noticing target constructions, the eye-movement indices revealed that the target unaccusative verbs were more likely to be attended to in the + complex condition than in the – one, as evidenced by the significantly greater number of fixations and longer fixation durations on target verbs. These results are in line with the findings from Study 2, in which increased task complexity had a significant positive impact on development in the knowledge of unaccusative verbs. Based on the findings from Study 2, it was speculated that the + complex tasks might have encouraged the participants to engage in more careful reading of the texts, resulting in repeated exposure and processing of target verbs. The eye-movement measures in the present study support this assumption, indicating that the target verbs did receive more attention from the participants in the + complex versions. The stimulated recalls offer further credence for this speculation, as the participants produced more comments related to noticing target unaccusative verbs as well as glosses related to the verbs.

Interestingly, however, task complexity did not affect the overall amount of attention paid to the glosses associated with the target verbs. That is, increased task complexity, according to the eye-movement data, did not encourage learners to check the glosses with greater frequency or process them longer. In fact, verb glosses were often ignored by participants; the average number of fixations to all verb glosses was below 4 for both texts regardless of task complexity level, although Text 1 and Text 2 included 8 and 7 target verbs respectively. Perhaps if participants were familiar with the prototypical meaning of the verb, they felt it unnecessary to check the verb glosses after they had visited a few of them, given that the grammatical information entailed in them

(i.e., unaccusative usage) was the same. In other words, the underlining of the target verbs might have been adequate to remind them of the target form-meaning mapping.

It was also found that task complexity had no significant impact on the noticing of pseudo-words and their glosses, as indicated by a lack of a significant difference in the number and sum of eye-fixations at pseudo-words and their glosses across the two task complexity conditions. One reason for this finding may be that the processing of pseudo-words was less essential to task completion than that of the unaccusative verbs. If the target pseudo-words had been selected on the basis of degree of task-essentialness, task complexity might have affected the extent to which they were attended to and processed.

CHAPTER 6

SUMMARY AND CONCLUSION

I. Summary of the Thesis

The present thesis has reported the results from three experimental studies that investigated whether and how manipulating task demands and glossing texts affect L2 reading and development in the knowledge of target constructions. Working memory capacity was included as an individual difference factor moderating the effects of task demands and glossing. Following a 2x2 experimental design with the two independent variables (i.e. task complexity and glossing), Studies 1 and 2 examined whether Korean undergraduate students' L2 reading comprehension was affected by task complexity and glossing, and if they acquired varying amounts of knowledge of the target constructions under different task conditions via pretest – posttest – delayed posttest. As the first two studies focused on off-line test scores, no evidence could be obtained regarding learners' internal processes while engaging in L2 reading tasks. As such, a third study, an eye-tracking project, was conducted in order to document learners' on-line reading processes. The eye-movement data, triangulated with stimulated recall protocols, offered insights into how task complexity influenced the learners' reading processes and noticing of glossed target linguistic constructions. The results collected from each study provided valuable information about the impact of manipulating task demands and glossing on L2 reading comprehension and L2 learning, and the role of working memory as a moderator of the effects of task complexity and glossing. In the following section, a more detailed summary of each study is presented.

1. Study 1

So far, TBLT studies have been confined to how task demands affect learners' linguistic production and learning of L2 constructions as a by-product of engaging in

interactive tasks, while the potential effects of task manipulations on L2 reading have been largely neglected. Additionally, research into the efficacy of glossing has predominantly focused on the acquisition of L2 vocabulary, while the potential usefulness of glossing in promoting L2 grammatical learning has been relatively unattended to. Thus, in order to address these gaps, Study 1 was conducted to examine whether the cognitive complexity of L2 reading tasks and the glossing of reading texts would influence L2 English reading comprehension and the learning of target L2 constructions contained in the reading passages.

The study employed a pretest – posttest – delayed posttest design with two treatment sessions. The target features were 17 English unaccusative verbs and ten pseudo-words. The participants were 52 Korean college students learning L2 English. They were randomly assigned to one of [– complex, – glossing], [+ complex, – glossing], [– complex, + glossing] and [+ complex, + glossing] conditions and read a text in each treatment session. Under the – complex condition, the task involved reading a text and answering multiple-choice reading comprehension questions as learners normally would when doing the reading section of a TOEFL test. Under the + complex condition, the paragraph order was jumbled so that the participants had to rearrange the paragraphs into a coherent order, in addition to answering reading comprehension questions. Under the + glossing condition, Korean definitions of the target constructions were provided in the margins of the texts, whereas no such information was provided in the – glossing condition. Reading comprehension was measured with 14 multiple-choice items for each text, and the learning of target constructions was assessed in terms of accuracy via a grammaticality judgment test and word form and meaning recognition tests. Working memory capacity, which is central to L2 reading comprehension and L2 learning, was included as a potential moderating variable. A forward digit span test and a nonword repetition test were used to measure phonological short-term memory, and a

backward digit span test and an operation span test were employed to assess complex working memory.

The results of mixed-effects modelling indicated that increased task complexity had significant negative effects on vocabulary form recognition scores in a delayed posttest. It was speculated that, in the + complex condition, the increased task complexity might have directed participants' attention to the paragraph-ordering task, and thus away from the pseudo-word forms. Glossing, by contrast, was shown to have a positive influence on the recognition of target pseudo-word forms and meanings. Interestingly, glossing had lasting effects on word meaning recognition, but only immediate effects on form recognition. In addition, it took longer for the participants to respond to word meaning recognition items and they were less confident in their responses, when compared to word form recognition items. This finding was interpreted as suggesting that word meanings were processed at a deeper level compared to word forms, resulting in more robust retention. It was found, however, that neither task complexity nor glossing affected reading comprehension scores or grammaticality judgment test scores. When it comes to the moderating role of working memory capacity, participants with higher NWS scores obtained higher reading comprehension scores for Text 1 in the [+ complex, + glossing] condition.

Several problems arose when analysing the data obtained from Study 1. First, task manipulation proved unsuccessful in terms of increasing task complexity. As mentioned earlier, a paragraph-ordering step was added to the + complex condition, and this was expected to increase the cognitive demands imposed on the participants. This manipulation, however, only affected the word form recognition scores, having a negative influence, and the participants' self-reports on the perceived level of mental effort were also not significantly different between the + and – complex conditions. Next, neither task complexity nor glossing had a significant impact on reading

comprehension scores. As discussed in Chapter 3, the mean scores were high, implying a ceiling effect, which could have lowered the likelihood of observing the influence of task complexity and glossing on reading comprehension scores. Thus, it was concluded that the difficulty of the reading comprehension tests might need to be adjusted in the main study. Last but not least, the inferrability of the target pseudo-words was not considered equivalent, which indicated the need to select target words more carefully.

2. Study 2

Study 2 replicated Study 1 but on a larger scale and with several modifications to the research methodology. Eighty-eight Korean undergraduate students were recruited, and they were randomly assigned to one of four 2x2 experimental conditions: [– complex, – glossing], [+ complex, – glossing], [– complex, + glossing], and [+ complex, + glossing]. The same reading texts were used, but the tasks were slightly different from those used in Study 1. Assuming that paragraph-level task manipulation was not effective in promoting learners' intensive linguistic processing of texts, Study 2 manipulated the reading tasks at a more localized level. More specifically, each paragraph of the texts was divided into two subparts under the – complex condition, but three to four subparts under the + complex condition. The participants were instructed to determine the correct order of the subparts before answering reading comprehension questions. In order to control the time spent on task completion and thereby detect the influence of task manipulation more clearly, 25 minutes was set as the time limit for task completion. Also, reading comprehension questions that reduced reliability in Study 1 were deleted, so that the discriminability of the participants' text understanding would be enhanced. In addition, some new target words were selected after discussing these with three Korean speakers, doctoral students in applied linguistics, to control the inferrability of words from the context.

The results revealed that, again, neither task complexity nor glossing had significant effects on reading comprehension scores. Increased task complexity, however, was shown to facilitate the learning of target unaccusative verbs in an immediate posttest. That is, re-arranging three to four subparts of each paragraph, instead of only two subparts, appeared to encourage closer inspection of a given text, and accordingly facilitated the processing of target verbs. Yet, increased task complexity was shown to have a negative impact on word meaning recognition in a delayed posttest. Probably, when reading the texts under the + complex condition, learners paid less attention to the target words and focused more on the text-ordering task. Additionally, the words might not have been essential for task completion, although they were carefully selected. Hence, the participants might have decided to prioritize the ordering task and reading comprehension questions over discovering word meanings (through either inferring from the context or referring to the glosses provided).

As in Study 1, glossing had positive effects on word meaning recognition scores, confirming the previous findings for the facilitative role of glossing in L2 lexical learning. However, glossing had a negative impact on word form recognition scores, which ran counter to the findings from Study 1. In Study 1, tasks were manipulated on a paragraph level, and the participants were allowed to stay on the task as long as necessary. Thus, participants could have had some surplus attentional resources even in the + complex condition, which might have allowed them to attend to the glossed word forms. By contrast, in Study 2 where task demands increased with the time limit, participants would have been under greater cognitive load and thus decided to take advantage of the glosses rather than paying extra attention to word forms.

With respect to the moderating effects of working memory capacity, participants with higher operation span scores achieved higher reading comprehension scores for Text 2 in the – complex and + glossing condition, implying the importance of complex

working memory for storing a gloss in short-term memory and utilizing it for textual processing. In other cases, however, working memory indices shared negative correlations with reading comprehension scores or learning scores, rendering it challenging to interpret the results. One possible explanation for the results is that the participants' cognitive maturity might have been at its peak (Craik & Bialystok, 2006) and/or their L2 English proficiency was, overall, above intermediate level, and thus it could have been difficult to observe any significant findings. More specifically, the participants' ages ranged from early to mid-twenties, and they were all sampled from a highly reputable university in Korea, which requires a certain level of English proficiency to be admitted. Indeed, the coefficients of variation (ratio of *SD* to Mean) for the working memory indices ranged from .04 to .12, indicating low variance, and effect sizes were overall small. That said, perhaps learners with a wider spectrum of cognitive ability, e.g. younger learners or learners with more diverse educational backgrounds, might present a clearer picture of how working memory capacity moderates the effects of task complexity or glossing on L2 reading and L2 learning.

3. Study 3

Motivated by the lack of evidence on learners' cognitive processes influenced by task manipulation, the third study was conducted using eye-tracking technology and stimulated recall protocols. The research questions asked whether learners' reading processes would differ when engaging in– versus + complex tasks and whether task complexity would affect the noticing of glossed target linguistic constructions. The participants were 38 Korean graduate students. They read the same texts that were used in Study 2 under – complex and/or + complex task conditions, while their eye-movements were recorded with an in-built eye-tracker. Eleven students were further invited to take part in stimulated recall protocols after completing both reading tasks.

The results indicated that the participants processed the texts more carefully and thoroughly under the + complex condition in comparison to the – complex condition. The participants' eye-fixations were longer and more frequent on the texts as well as on the task as a whole under the + complex condition. Additionally, the numbers of forward saccades and regressions were significantly greater in the + complex condition, implying the recursive processing of texts. In similar vein, stimulated recalls revealed that the participants perceived the + complex versions as more difficult and they felt less confident about their performance under the + complex conditions, compared to the – complex versions. They also reported that they employed a greater variety of reading strategies and took advantage of more diverse clues when performing the + complex versions. Eye-movement analysis further demonstrated that the target verbs received significantly more eye-fixations, which supports the findings obtained from Study 2, in which increased level of task complexity facilitated development in the knowledge of target unaccusative verbs. Different task conditions, however, had only limited influence on the eye-movement measures for pseudo-words and glosses, suggesting a lack of interaction between task complexity and glossing.

II. Overall Discussion

1. Impact of task complexity on L2 reading tasks

The present thesis has revealed that task demands might affect learners' reading processes as well as development in knowledge of the target linguistic constructions contained in the texts, either positively or negatively. The findings fit neatly into Khalifa and Weir's (2009) *cognitive processing model for reading comprehension* (for a schematic diagram, see Figure 2). Within this model, task manipulation can come into play through the *goal setter*, which produces different permutations of the two dimensions of reading (careful vs expeditious and local vs global), which in turn affects

the entire reading process. When applied to Study 1, the additional paragraph-ordering step in the + complex versions might have led the participants to read the texts more carefully in order to confirm the main idea of each paragraph and sequence them logically. In Study 2, the sentence-rearranging task in the + complex versions could have necessitated more intensive textual processing to arrive at an accurate understanding of each sentence. In other words, the + complex reading tasks in Studies 1 and 2 appear to have put more emphasis on the importance of careful reading compared to the – complex counterparts. Following this line of logic, when the + complex reading tasks used in Studies 1 and 2 are compared, the latter might have been more complex than the former due to the different depth of reading required for successful task completion. While those in Study 1 could be accomplished by conceptually organizing global (paragraph-level) ideas, Study 2 had to be carried out by processing the texts attentively and thoroughly. As demonstrated in the present thesis, it seems viable to adjust the level of cognitive demands imposed on learners by manipulating various dimensions of an L2 reading task.

First, reading tasks may be manipulated in terms of the extent to which careful reading is required. If a reading task necessitates thorough and scrupulous processing of textual information, the cognitive demands may be greater compared to its counterpart that can be completed with shallow and superficial processing (Craik & Lockhart, 1972; Craik & Tulving, 1975). For example, at a local level, reading a list of sentences to fill in blanks (cloze task) may require more careful local reading than reading the same text for pleasure. Also, at a global level, reading to prepare for a discussion (e.g. Taillefer, 1996), reading for coherence (e.g. Horiba, 2000), reading for critique (e.g. Horiba, 2013) and reading for memorization or retelling (e.g. Yoshimura, 2006) may be cognitively more demanding than reading the same text freely. In addition, at either a local or global level, careful reading may entail attentive and intensive linguistic processing of textual

information (Nassaji, 2003, 2007, 2014), and thereby provide learners with more opportunities for exposure to new TL features, noticing gaps and establishing novel TL form-meaning mappings. Robinson (2011) also claims that more complex tasks might promote heightened attention to and greater depth of processing of the input provided, resulting in longer-term retention than in simpler tasks. In that sense, increased task demands with regard to the *depth* of reading may facilitate L2 development, which resembles Robinson's task manipulation along the *resource-directing dimension* or Bialystok's (1994) *analysis* of linguistic samples of restructuring interlanguage (see Figure 20).

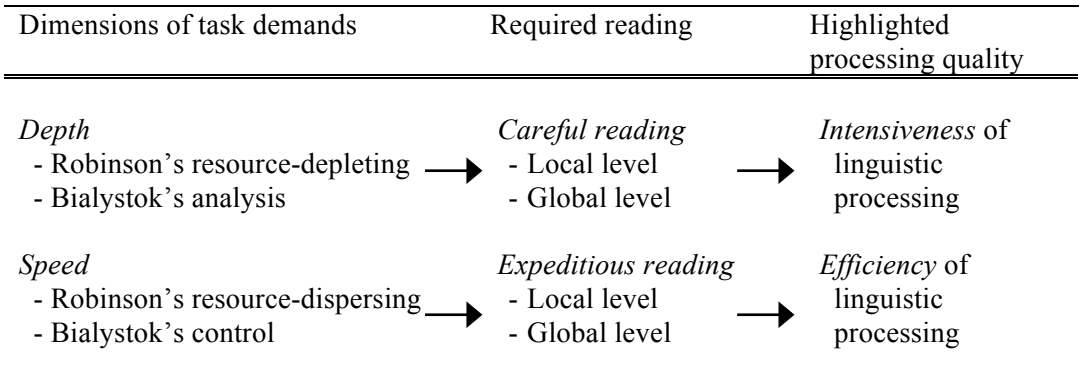

Figure 20. Proposed relationship between task demands and reading process

Although expeditious reading has been relatively unattended to by researchers, reading tasks can also become cognitively demanding when they need to be completed quickly. Expeditious reading requires selective, efficient and automatic lower-level text processing (Birch, 2007; Nassaji, 2014), which may pose even greater problems for many L2 readers who suffer from incompetent decoding proficiency (Khalifa & Weir, 2009). For example, identifying false information from a list of sentences or extracting the main idea of a text within a time limit can feel more demanding for L2 readers, in comparison to performing the same tasks without such time pressure. In this regard, the reading tasks used in Study 2 could have been more demanding than those in Study 1, as Study 2 involved a time limit. It can be further hypothesized that expeditious reading, at either local or global level, may provide learners with opportunities to practise

decoding skills, such as word recognition and syntactic parsing. In other words, increasing task demands by setting a time limit may impose performative demands on learners and thereby help them to automatize and proceduralize lower-level processing. In that sense, the manipulation of reading tasks in terms of the required *speed* of reading may be seen as equivalent to Robinson's *resource-dispersing dimension* or Bialystok's concept of *control* over existing linguistic knowledge.

Lastly, when the depth and speed of reading remain constant, global reading may impose greater processing demands on readers compared to local reading, due to a larger amount of textual input to be comprehended. In comparison to local reading that can be accomplished with word recognition and syntactic parsing, global reading additionally necessitates connecting propositional units and creating a coherent textual representation. Thus, reading tasks that include understanding extended textual input, such as a paragraph or a multi-paragraph passage, will be more demanding than tasks that can be completed by comprehending a limited amount of input, such as words or sentences. The findings from Brunfaut and McCray's (2015) research support this assumption. This study found that processing demands increased from CEFR level A1 to B2, indicating that a wider scope of reading was more demanding than processing a limited amount of input. In the case of the present thesis, the scope of reading was identical between the + and – complex versions in both Studies 1 and 2. In future research, investigating how task manipulation in terms of the depth, speed or scope of reading affect learners' reading processes and noticing and/or learning of target constructions could generate insightful findings that will be valuable for refining a theoretical and pedagogical framework for a task-based approach to L2 reading.

2. Glossing in L2 reading and L2 learning

The fundamental rationale for glossing texts for L2 readers is that glosses can ease the initial construction of micro-structures by means of assisting meaning retrieval, and

thereby promote text understanding. Previous findings, however, have had mixed findings regarding the efficacy of glossing on L2 reading comprehension, and the present thesis has also failed to yield significant results. One possible explanation is that the glossed items might not have been essential for answering the reading comprehension questions (Loschky & Bley-Vroman, 1993). If the target items were selected on the basis of the degree of task-essentialness, the effects of glossing might have surfaced in the reading comprehension scores. In addition, reading comprehension is by no means a unitary construct, rather it subsumes multi-layered mental representations of a given text, most notably, a local-level representation built through lower-level processes and a global-level representation based on higher-level processes (Khalifa & Weir, 2009; Kintsch, 1998; Perfetti, 1999). While glossing was originally purported to function at a local level through facilitating semantic access to unknown linguistic items, glossing has also been shown to permeate through inferential comprehension, presumably by allowing L2 readers to spare attentional resources that might otherwise be used for decoding (e.g. Ko, 2005). Thus, in future studies, it will be desirable to develop reading comprehension items that necessitate processing glossed items, while at the same time accounting for different levels of text understanding, so that whether and how glossing functions in the process of reading comprehension can be elucidated more clearly.

Next, including the results of this thesis, glossing, in general, has been found to promote L2 lexical learning. Additionally, glossing has been shown to exert a longer-term influence on promoting word meaning recognition (i.e. significant positive effects on delayed posttests) compared to form recognition scores (i.e. significant effects only on immediate posttests). The *Involvement Load Theory* was called upon to explain the differential efficacy of glossing on word form versus meaning recognition scores. Hulstijn and Laufer (2001), drawing on Craik and Lockhart's (1972) concept of *depth of*

processing, postulate that processing the meaning of a new lexical item might take place at a rather deep level, but that of form at a shallow level. They further predict that, accordingly, retention of the semantic encoding of a new word will be more robust than registering a word's phonological form, which is supported by the findings of Study 1.

When it comes to word form recognition, glossing had a positive impact in Study 1, but a negative influence in Study 2, in which the tasks became relatively more demanding in comparison to the ones used in Study 1. The contrasting findings between Studies 1 and 2 indicate that the effects of glossing on word form recognition may interact with the level of cognitive task demands. When learners can resort to some residual attentional resources during reading, as in Study 1, glossing may draw learners' attention to the forms of glossed items, especially considering that glossed items are usually underlined or numbered in order to inform the reader that the meanings of items are provided. However, when learners are under increased task demands, they may look directly at glosses rather than attending to word forms. More empirical research into the joint influence of task complexity and glossing on word form and meaning recognition may provide valuable evidence for a fuller understanding of how to promote L2 lexical learning through glossing.

Last but not least, glossing is normally done for isolated linguistic items and is expected to facilitate mapping of form and meaning of individual items. Hence, glossing has generally been considered more suitable for promoting lexical learning. Previously, Guidi (2009) and Martinez-Fernández (2010) investigated if L1 glosses could facilitate the learning of L2 Spanish grammatical constructions, which were pioneering attempts to utilize glosses for promoting the learning of L2 grammatical features. Yet, the target grammatical constructions in these studies, which were the Spanish present perfect, impersonal *se* and subjunctive, entailed complex grammatical conjugations, which necessitated abstracting the underlying system-wide rules from

individual glossed cases. Thus, in order to facilitate the acquisition of these grammatical constructions, glossing might have needed to be provided for longer, with wider cases, so that learners could generalize the overarching rule and apply it to new cases.

Interestingly, although significance was not quite achieved in Study 1 ($z = 1.85$, $p = .06$, $C = .82$), the thesis has demonstrated the potential efficacy of glossing to promote the learning of English unaccusative verbs. One of the factors relating to the learnability problem of unaccusative verbs is that meaning cannot be simply drawn from the prototypical meaning of a verb (e.g. *I collect stamps*), but additionally from the verb argument construction in a particular context (e.g. *Gas and oil can collect in sandy layers*). As such, in order to acquire the semantic restrictions of unaccusative verbs (i.e. the need to switch the agent with the object in subject position), exposure to grammatical usages of individual unaccusative verbs is critical (Goldberg, 1998; Ono & Budwig, 2005; Zyzik, 2009). That being the case, glossing might be more useful for facilitating the learning of constructions that exhibit lexical patterns and that are acquired in an item-based manner. Examples of such constructions include, in addition to unaccusative verbs, ditransitive verbs, multi-word verbs (e.g. phrasal verbs), collocations and formulaic expressions, to name but a few. Indeed, L2 learners have been reported to struggle with acquiring these constructions, typically due to low input frequency and unique distribution. Acquisition of these features requires item-by-item and lexically-specific exposure to individual usages, which resembles that of lexical learning, and thus might be more susceptible to glossing. In that sense, the selection of target constructions for glossing might be guided by theories that emphasize input frequency and pattern analysis in language acquisition, such as systemic functional grammar (Halliday & Matthiessen, 2013), the usage-based theory of language learning (Tomasello, 2003), frequency-based accounts of SLA (N. Ellis, 2012), the competition

model of language learning (MacWinney, 2012) or a corpus-based approach to language studies.

3. Moderating effects of working memory capacity

Despite the long-attested centrality of working memory capacity in reading comprehension and language learning, the present thesis failed to detect a significant moderating effect of working memory on either reading comprehension scores or learning target constructions. In Study 2, in particular, some unexpected inverse relationships were found, rendering the results even more puzzling. In a recent study conducted by Serafini and Sanz (2015), though, similar findings were reported. In their study, the contributions made by complex working memory and phonological memory to the learning of ten Spanish grammatical structures were assessed over the course of developmental stages, namely, beginning, intermediate and advanced L2 Spanish proficiency. The results of longitudinal assessment of the relationships between working memory indices and knowledge of target constructions revealed that beginners and intermediate learners were shown to rely on their working memory, especially phonological memory, to a significant extent for L2 learning. In contrast, for advanced learners with much exposure to and experience in Spanish, working memory scores shared negative relationships with posttest scores. This result was consistent with previous research findings that showed a lack of influence of working memory capacity on L2 development (e.g. Coughlin & Tremblay, 2013; Foote, 2011; Hummel, 2009). Based on these findings, they suggested that the role of cognitive ability, such as working memory capacity, might decrease as learners receive increasingly more L2 exposure. In addition, they further speculated that, for advanced learners, working memory measures could have yielded significant findings if the task had been more challenging (e.g. spontaneous oral production tasks).

In the field of TBLT, working memory has been shown to play only a marginal role in moderating the effects of task complexity. In Baralt's (2010) study, diverse span scores (operation span, counting span and reading span) correlated negatively with L2 development scores in a computer-mediated condition. Also, in Kormos and Trebits' (2011) study, participants with the lowest and highest digit span scores achieved the lowest mean values in terms of the ratio of subordinate clauses, whereas those in the middle range obtained the highest values. Together with the results from the present thesis, these findings seem to indicate that, as Kormos and Trebits claim, the relationship between working memory and learners' task performance or L2 learning might not be linear. That said, as presented in Figure 21, the moderating effects of working memory capacity may be evidenced between the lower threshold and the upper threshold of task complexity as well as learners' cognitive and L2 ability. To be more specific, when the task is too easy or too demanding for learners, their linguistic performances or L2 development from engaging in the task may become comparable, rendering it challenging to observe any moderating role of working memory capacity. Likewise, if learners' cognitive or linguistic ability is too low or too advanced, it will be difficult to obtain sufficient variance among learners, which is essential to explore the role of working memory capacity as a potential individual difference factor in language learning.

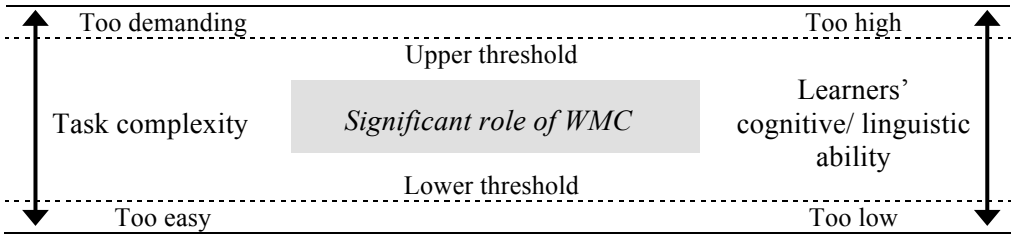

Figure 21. Significant role of WMC between upper and lower thresholds

III. Implications

1. Theoretical implications

As reflected in the above discussion, the findings of this thesis have several theoretical implications. First, in this thesis, task complexity affected learners' reading processes as well as their knowledge of target constructions contained in the text. Yet, previous models of task complexity, such as Skehan's *Limited Capacity Model* (Skehan, 1998, 2009; Skehan & Foster, 2001) and Robinson's *Cognition Hypothesis* (Robinson, 1995b, 2001a, 2011), do not make predictions with respect to how cognitive task demands will affect learners' reading behaviour and L2 development. In order to fill this gap, in this thesis, a rough framework of L2 reading task complexity was proposed based on previous empirical findings and drawing on Khalifa and Weir's (2009) cognitive processing model for reading comprehension. It was suggested that task complexity could be manipulated in terms of the depth, speed and scope of reading required. According to this framework, a task that requires careful reading of a long article within a specified time limit is considered more complex than a task that can be accomplished through casual reading of a list of words or sentences without any time limit. It was further hypothesized that increasing task complexity by manipulating the depth of reading required would have an impact on learners' analysis of new L2 constructions by encouraging careful and intensive linguistic processing of the text, whereas manipulating the required speed of reading might influence learners' control over their existing L2 knowledge through facilitating the automatisation of decoding skills. In order to test the applicability of this framework, more empirical investigations are imperative so that we can deepen our understanding of how the cognitive demands of L2 reading tasks affect learners' L2 reading processes, comprehension outcomes and noticing and/or acquisition of target linguistic constructions by engaging in L2 reading tasks.

Interestingly, in this thesis, increased task complexity had long-term negative effects on word recognition scores, in contrast to its positive influence on learning target unaccusative verbs. More specifically, greater task demands had a negative impact on delayed word form recognition scores in Study 1 and delayed word meaning recognition scores in Study 2. It should be mentioned that, although the target lexical items were carefully selected, they could have been not critical for successful task completion. Provided that, the findings of this thesis may imply that the increased demands of L2 reading tasks can strengthen learners' tendency to allocate their attentional resources selectively, by prioritizing the part of input that is necessary for task completion. To put it another way, as task complexity increases, task-essential constructions may receive heightened attention from learners, whereas constructions that are not crucial for task completion will be more likely to be skipped by learners. Whether this propensity for strategic and selective processing of textual input is indeed reinforced for cognitively demanding L2 reading tasks should be tested in future research.

The findings of this thesis also indicate that the efficacy of glossing might interact with the cognitive complexity of the reading task. As reviewed earlier, the theoretical rationale for glossing is that glossed items will not only improve learners' L2 reading comprehension by facilitating meaning retrieval at the decoding stage, but also promote establishment of the form-meaning connections of glossed items. The findings of the present thesis indicate that this assumption is premised on the condition that learners have residual attentional capacity after completing a reading task. As discussed in the previous section, glossing facilitated noticing of both the forms and meanings of target words in Study 1, in which no time limit was set. By contrast, when tasks became more demanding in Study 2, glossing hindered learners' registering of target word forms. In other words, as the reading task becomes more demanding, learners appear to be less able to attend to the forms of glossed items but are instead motivated to resort to glosses

in order to save their attentional resources. Hence, future research into glossing may need to take the processing demands of reading tasks into account and investigate the distinctive efficacy of glossing on acquisition of the forms and meanings of target linguistic items.

2. Methodological implications

The present thesis has a number of methodological implications. In this thesis, learning assessments (i.e. grammaticality judgment tests and vocabulary recognition tests) were constructed using the research software E-Prime, which allowed the efficient randomization of test items as well as the measurement of learning from different perspectives. First, although the same test items were used over pretest, posttest and delayed posttest, items were randomized within each test so that practice effects and ordering effects could be reduced. Next, learning could be assessed from various angles, including the accuracy of responses, reaction times, binary confidence ratings and source attributions (Rebuschat, 2013; Rogers, 2016). Data from these different sources allowed data triangulation and thereby provided a fuller understanding of the nature of acquired knowledge, if any. Thus, in future research, it is recommended to gather multiple sources of data, which will generate a more nuanced understanding of the relationships among variables.

In addition, in the third study, eye-movement data were combined with stimulated recall protocols, which provided richer and more solid findings. To be more specific, eye-movement data offered insights into lower-level reading processes, whereas stimulated recall protocols generated insights into learners' higher-level processes (Bax, 2013; Brunfaut & McCray, 2015). Thus, these two sources of data, when combined, painted a more accurate picture of how task complexity affected learners' reading behaviours and noticing of glossed target linguistic constructions. Moreover, the data allowed validation of the construct of task complexity. More specifically, the eye-

movement data revealed that when reading texts under the + complex condition, participants engaged in more thorough and recursive processing of them. The stimulated recall protocols showed that the participants perceived the + complex versions to be more demanding than the – complex versions. That said, in future research, eye-tracking technology, followed by stimulated recalls, may provide researchers with valuable insights into how reading tasks involving different features affect learners' reading processes in distinctive ways.

Last but not least, the data gathered in the present thesis were analysed with mixed-effects modelling using the statistical program R, which could account for the random variability across and within participants and items. In a standard regression analysis, data analysis is conducted on summed and averaged values. In this case, idiosyncrasies nested in participants or items could not be accounted for. Indeed, when the data in Study 1 were analysed using repeated measure ANOVAs, glossing emerged as a significant factor promoting the learning of target unaccusative verbs ($F(2, 96) = 5.57$, $p = .01$, partial $\eta^2 = .10$). However, this significance disappeared when the same data were analysed using mixed-effects modelling ($z = 1.85$, $p = .06$, $C = .82$). The different results seem to suggest that the significant relationships found with the ANOVAs in Study 1 might be partially attributable to random variances caused by participants and/or items. Thus, for empirical studies in an SLA setting that typically involve multiple participant-level as well as stimulus-level independent variables, mixed-effects modelling will serve as a powerful statistical tool, allowing wider generalization of research findings.

3. Pedagogical implications

The present thesis included both task-based manipulation (i.e. task demands) and text-based modification (i.e. glossing), and examined their efficacy in promoting L2 reading comprehension and the acquisition of L2 constructions. While more empirical

evidence should be accumulated for any pedagogical implications to be drawn, the findings of the present thesis can be used as a basis for tentative pedagogical implications. As discussed earlier in the thesis, reading comprehension can be achieved mainly through conceptual processing (Sharwood Smith, 1986), and learners have a natural tendency to prioritize meaning extraction at minimum attentional cost (VanPatten, 2012). Thus, L2 learners should be helped to reallocate their attentional resources to the processing of L2 linguistic features for any L2 development to occur. However, current trends in L2 reading instruction have followed diverging paths, i.e. overemphasis on either meaning comprehension or explicit L2 learning using L2 texts. The findings of this study suggest that L2 reading tasks that encourage learners to read a given text carefully and thoroughly to achieve a meaningful objective might be useful in promoting the development of L2 proficiency through L2 reading, without interrupting L2 reading comprehension.

With respect to glossing, research findings so far, including the present thesis, have shown the positive effects of glossing on L2 lexical learning. In addition, the importance of task-essentialness has emerged, suggesting that when there is no need to infer the meaning of a word, its form or meaning might only be processed superficially. That said, if a word's meaning should be processed for successful task completion, but when the context around the word does not provide enough clues for meaning inference, glossing might be able to drive a learner's attention to its meaning. However, it was also found that if a reading task is too demanding, forms are less likely to be processed when glossed. This finding suggests that pre- or post-reading activities can be useful for boosting the efficacy of glossing by providing learners with additional opportunities to reinforce the memory traces of glossed words, both form and meaning. Laufer (2009), based on a synthetic review of empirical studies on incidental L2 lexical learning, also

concluded that "input together with engaging word-focused activities and frequent rehearsals" (pp. 341–342) is most likely to produce positive results.

IV. Limitations and Future Directions

There are several limitations to the present thesis. First of all, the reading tasks in this thesis involved reading the passages provided while answering multiple-choice reading comprehension questions, which is fundamentally what learners would normally do when taking a language aptitude or proficiency test. In this regard, there may be concerns regarding the ecological validity (the extent to which the methods, materials, and setting of a study approximate target language use settings) of the task in terms of its resemblance to real-world reading tasks. For instance, when compared to tasks such as reading the manual to use a new electronic device, reading news articles or editorials, or even flipping through a menu at a restaurant, the task used in this thesis may appear to have low ecological validity. Yet, the applicability of the reading tasks used in this thesis has implications in many cases in academic settings where learners take exams. Also, in the present thesis, the extent to which the reading task invoked the kind of cognitive processes that are essential in performing a real-world task was considered more important than how much the task approximated to a target task in its appearance. For instance, if a learner can read a given text and identify the author's intention, as one of the reading comprehension questions required, we can make a valid assumption that the learner is likely to perform other similar tasks, such as reading an editorial and understanding the author's opinion.

Some tension between ecological validity and construct validity (the degree to which a test measures what it purports to measure) also arose when reconstructing the reading tasks for the eye-tracking project. The primary purpose of Study 3 was to reveal the underlying cognitive processes when performing the + and – complex reading tasks

used in Study 2. Hence, ideally, the original task layout needed to be maintained as much as possible in order to replicate the kind of cognitive processes enacted in Study 2. On the other hand, the accurate capture of learners' eye-movements required inevitable reformatting of the task layout, such as using different font and line spacing, a changed paragraph structure and forced sequential task completion, among others. With respect to this methodological dilemma posed in eye-tracking research, Spinner et al. (2013) suggest that ecological validity might receive emphasis when the research question is to explore learners' internal processes in normal reading, whereas manipulating textual prompts would be necessary when investigating learners' cognitive processing (e.g. noticing) of small grammatical functors. Given that Study 3 tapped into both reading processes and the noticing of target constructions, the researcher had to find some middle ground between ecological validity (i.e. exact replication of the task format of Study 2) and construct validity (i.e. accurate recording of eye-movements influenced by task demand manipulation).

Another limitation was the low discriminability of the reading comprehension questions employed in the present thesis. The items were taken from previously administered TOEFL tests, along with the two reading passages, in an attempt to increase the reliability and validity of the tests. Despite that, the reliability of the tests was shown to be only minimal (ranging from .14 to .57), supposedly being partially responsible for the non-significant influence of task manipulation and glossing on the reading comprehension scores. As discussed earlier, the participants were, overall, homogeneous in terms of their English proficiency, which might explain the small variance in the data. In addition, in the case of Study 2 and the eye-tracking study, only one or two items followed each paragraph, posing inherent limitations in assessing learners' reading comprehension sensitively. It seems also noteworthy that previous studies have demonstrated the need to take into account the multi-dimensional nature of

reading comprehension, such as local and global comprehension (Kintsch & van Dijk, 1978) and text model and situation model (Kintsch, 1988). Therefore, it is imperative that future research designs reading comprehension tests more carefully for a fuller assessment by taking into account various aspect of text understanding and, in so doing, enhances the likelihood of maximizing variance in the data and examining the clear effects of independent variables on different types of reading comprehension.

It should be also noted that the English proficiency test (Reading and Language Use section of CPE) used in this thesis was consistently revealed to be highly demanding for the participants, presumably resulting in a floor effect. Given the mismatch between the participants' English proficiency and the target level of CPE, the results on the equivalence among the participants in terms of their English ability may need to be interpreted with caution. In future research, the proficiency test for measuring participants' target language level will need to be carefully selected.

In addition, the ten target words that were substituted by pseudo-words were selected based on two criteria, i.e. (a) the word appeared only once in the text and (b) the word was a noun. Additionally, the length of each pseudo-word was controlled, which was also important for the eye-tracking methodology. Yet, in future studies, it might be necessary to control for additional factors such as (c) the word is concrete or abstract, (d) the word is inferable or non-inferable from the context, (e) the word is essential for task completion, and (f) the position of the word in a sentence or paragraph is identical (e.g. initial vs middle vs final). While Korean speakers majoring in applied linguistics were invited to assist with careful selection of the target words, all of the conditions from (a) to (f) could not be met simultaneously. For stronger internal and external validity of research findings, attention will need to be paid in future studies to control for most of these conditions.

Also, in the eye-tracking study, mainly due to practical restrictions such as the small number of participants and the limited time allowed for data collection, the effects of glossing and the moderating effects of working memory capacity on L2 reading processes could not be explored. Had this been done, it would have been possible to examine (a) whether glossed items did receive heightened attention compared to unglossed counterparts and (b) whether learners with higher working memory measures exhibited different reading behaviours when compared to those with lower working memory measures. In addition, a repeated measurement design (i.e. pretest – posttest – delayed posttest) could not be employed, and thus how differential reading processes resulted in different learning outcomes could not be explored. That said, replication of the eye-tracking study in this thesis with more participants through longitudinal data collection may paint a more complete picture of the effects of task complexity and glossing on L2 reading processes and their implications for L2 learning. It should be further noted that the sampling rate of the eye-tracker used in Study 3 was low for capturing eye-fixations and durations made on individual target constructions or the related glosses, and hence the results of the study may need to be interpreted with caution.

Despite these limitations, the present thesis has shed some light on how task complexity affects learners' L2 reading and L2 development, which has long been neglected in the fields of both task-based language teaching and L2 reading research. More empirical research on this topic can provide valuable information for predicting and explaining the effects of cognitive task demands on L2 reading. In addition, the alleged role of glossing in L2 reading, i.e. facilitating reading comprehension and L2 learning, was shown to depend on task features, such as the cognitive demands put on learners. Hence, more studies on how the effects of glossing interact with various task features might provide meaningful findings on whether and how to use glosses and

under what task conditions. Lastly, given the centrality that working memory occupies in the field of both L2 reading and L2 learning, further empirical exploration is necessary to spell out how differential working memory capacity, as an individual difference factor, moderates learners' ability to cope with varying amounts of task complexity and processing of glosses contained in a text.

REFERENCE LIST

Adams, R. (2003). L2 output, reformulation, and noticing: Implications for IL development. *Language Teaching Research, 7*(3), 247-276.

Ahmadian, M. J., & Tavakoli, M. (2011). The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research, 15*(1), 35-59.

Al-Seghayer, K. (2001). The effects of multimedia annotation modes on L2 vocabulary acquisition: A comparative study. *Language Learning & Technology, 5*(1), 202-232.

Alanen, R. (1995). Input enhancement and rule presentation in second language acquisition. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 259-302). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.

Albert, Á. (2011). When individual differences come into play: The effect of learner creativity on simple and complex task performance. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (pp. 239-265). Amsterdam, The Netherlands: John Benjamins.

Alderson, J. C. (1984). Reading in a foreign language: A reading or a language problem? In J. C. Alderson & A. H. Urquhart (Eds.), *Reading in a foreign language* (pp. 1-24). London, UK: Longman.

Alderson, J. C. (2000). *Assessing reading*. New York: Cambridge University Press.

Allen, I. E., & Seaman, C. A. (2007). Statistics roundtable: Likert scales and data analyses. *Quality Progress, 40*(7), 64-65.

Allport, A. (1988). What concept of consciousness? In A. J. Marcel & E. Bisiach (Eds.), *Consciousness in contemporary science* (pp. 159-182). London, UK: Clarendon Press.

Alptekin, C., & Erçetin, G. (2009). Assessing the relationship of working memory to L2 reading: Does the nature of comprehension process and reading span task make a difference? *System, 37*, 627-639.

Alptekin, C., & Erçetin, G. (2011). Effects of working memory capacity and content familiarity on literal and inferential comprehension in L2 reading. *TESOL Quarterly, 45*(2), 235-266.

Alptekin, C., & Erçetin, G. (2015). Eye movements in reading span tasks to working memory functions and second language reading. *Eurasian Journal of Applied Linguistics, 1*(2), 35-56.

Alptekin, C., Erçetin, G., & Özemir, O. (2014). Effects of variations in reading span task design on the relationship between working memory capacity and second language reading. *The Modern Language Journal, 98*(2), 536-552.

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction, 16*(5), 389-400.

Baayen, R. H. (2008). Analyzing linguistic data: A practical introduction to statistics. Cambridge, UK: Cambridge University Press.

Baddeley, A. D. (1966a). Short-term memory for word sequences as a function of acoustic, semantic, and formal similarity. *Quarterly Journal of Experimental Psychology, 18,* 362-365.

Baddeley, A. D. (1966b). The influence of acoustic and semantic similarity on long-term memory for word sequences. *Quarterly Journal of Experimental Psychology, 18,* 302-309.

Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences, 4*(11), 417-423.

Baddeley, A. D. (2003a). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience, 4*, 829-839.

Baddeley, A. D. (2003b). Working memory and language: an overview. *Journal of Communication Disorders, 36*, 189-208.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47-89). San Diego, CA: Academic Press.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior, 14*, 575-589.

Balcom, P. (1997). Why is this happened? Passive morphology and unaccusativity. *Second Language Research, 13*(1), 1-9.

Baralt, M. (2010). *Task complexity, the Cognition Hypothesis, and interaction in CMC and FTF environments.* Unpublished Ph.D dissertation, Department of Spanish and Applied Linguistics, Georgetown University, Washington D.C.

Baralt, M. (2013). The impact of cognitive complexity on feedback efficacy during online versus face-to-face interactive tasks. *Studies in Second Language Acquisition, 35,* 689-725.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmnatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255-278.

Barton, K. (2015) *MuMIn: Multi-Model Inference. R package version 1.13.4.* http://cran.r-project.org/package=MuMIn.

Bates, D.M., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999999-0.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing, 30*(4), 441-465.

Bax, S., & Weir, C. (2012). Investigating learners' cognitive reading processes during a computer-based CAE reading test. *University of Cambridge ESOL Examinations Research Notes, 47*, 3-14.

Beck, I. L., McKeown, M. G., & McCaslin, E. S. (1983). Vocabulary development: All contexts are not created equal. *The Elementary School Journal, 83,* 177-181.

Bell, F., & LeBlanc, L. (2000). The language of glosses in L2 reading on computer: Learners' preferences. *Hispania, 83*, 274-285.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics. Identifying influential data and sources of collinearity*. New York, NY: Wiley.

Bernhardt, E. B. (2005). Progress and procrastination in second language reading, *ARAL, 25,* 133-150.

Bialystok, E. (1994). Analysis and control in the development of second language proficiency. *Studies in Second Language Acquisition, 16*, 157-168.

Birch, B. M. (2007). *English L2 reading: Getting to the bottom*. Mahwah, NJ: Lawrence Erlbaum.

Blau, E. K. (1982). The effect of syntax on readability for ESL students in Puerto Rico. *TESOL Quarterly, 16*(4), 517-528.

Bley-Vroman, R., & Masterson, D. (1989). Reaction time as a supplement to grammaticality judgments in the investigation of second language learners' competence. *University of Hawai'i Working Papers in ESL, 8*(2), 207-237.

Block, R. A., Hancock, P. A., & Zakay, D. (2010). How cognitive load affects duration judgments: A meta-analytic review. *Acta Psychologica, 134*, 330-343.

Blom, E., Paradis, J., & Sorenson Duncan, T. (2012). Effects of input properties, vocabulary size, and L1 on the development of third person singular –s in child L2 English. *Language Learning, 62*(3), 965-994.

Bowles, M. (2003). The effects of textual input enhancement on language learning: An online/offline study of fourth-semester Spanish students. In. P. Kempschinski & P. Pineros (Eds.), *Thoery, practice, and acquisition: Papers from the 6th Hispanic Linguistic Symposium and the 5th Conference on the Acquisition of Spanish & Portuguese* (pp. 359-411). Somerville, MA: Cascadilla Press.

Bowles, M. (2004). L2 glossing: To CALL or not CALL. *Hispania*, *87*(3), 541-552.

Bowles, M. (2008). Task type and reactivity of verbal reports in SLA. *Studies in Second Language Acquisition, 30*, 359-387.

Bowles, M. (2010). Concurrent verbal reports in second language acquisition research. *Annual Review of Applied Linguistics, 30*, 111-127.

Bowles, M., & Leow, R. P. (2005). Reactivity and type of verbal report in SLA research methodology: Expanding the scope of investigation. *Studies in Second Language Acquisition, 27*(3), 415-440.

Breen, M. P. (1987). Learner contributions to task design. In C. N. Candlin & D. Murphy (Eds.), *Language learning tasks. Lancaster practical papers in English language education*, Volume 7 (pp. 23-46). Englewood Cliffs, NJ: Prentice-Hall International.

Breen, M. (1989). The evaluation cycle for language learning tasks. In R. K. Johnson (Ed.), *The second language curriculum* (pp. 187-206). Cambridge, UK: Cambridge University Press.

Brindley, G. (1989). *Assessing achievement in the learner-centered curriculum*. Sydney, Australia: National Centre for English Language Teaching and Research.

British Council (2014). *Aptis*. Retrieved from http://www.britishcouncil.org/aptis.

Brunfaut, T., & McCray, G. (2015). Looking into test-takers' cognitive processes whilst completing reading tasks: a mixed-method eye-tracking and stimulated recall study. *ARAGs Research Reports Online*, AR/2015/001. London, UK: British Council.

Brunfaut, T., & Révész, A. (2015). The role of task and listener characteristics in second language listening. *TESOL Quarterly, 49*(1), 141-168.

Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 53-61.

Brünken, R., Plass, J. L., & Leutner, D. (2004). Assessment of cognitive load in multimedia learning with dual-task methodology: Auditory load and modality effects. *Instructional Science, 32*, 115-132.

Brünken, R., Steinbacher, S., Plass, J. L., & Leutner, D. (2002). Assessment of cognitive load in multimedia learning using dual-task methodology. *Experimental Psychology, 49*, 109-119.

Bygate, M., Skehan, P., & Swain, M. (2001). *Researching pedagogic tasks, second language learning, teaching and testing*. Harlow: Longman.

Candlin, C. (1987). Towards task-based language learning. In C. Cadlin and D. Murphy (Eds.). *Language learning tasks*. Englewood Cliffs, NJ: Prentice Hall.

Case, R., Kurland, M. D., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology, 33*, 386-404.

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Boston, MA: Houghton Mifflin.

Chaudron, C. (1985). Intake: On models and methods for discovering learners' processing of input. *Studies in Second Language Acquisition, 7*, 1-14.

Cheung, H. (1996). Non-word span as a unique predictor of second-language vocabulary learning. *Developmental Psychology, 32*, 867-873.

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *Modern Language Journal, 80*(2), 183-198.

Chung, T. (2014). Multiple factors in the L2 acquisition of English unaccusative verbs. *IRAL, 52*(1), 59-87.

Cierniak, G., Scheiter, K., & Gerjets, P. (2009). Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Computers in Human Behavior, 25*, 315-324.

Cohen, A. D. (1994). *Assessing language ability in the classroom*. Boston, MA: Newbury House/Heinle and Heinle.

Cohen, A. D., & Upton, T. A. (2006). *Strategies in responding to the New TOEFL reading tasks* (TOEFL Monograph Series Report No. 33). Princeton, NJ: Educational Testing Service. http://www.ets.org/Media/Research/pdf/RR-06-06.pdf.

Cohen, A. D., & Upton, T. A. (2007). 'I want to go back to the text': Response strategies on the reading subtest of the new TOEFL®. *Language Testing, 24*(2), 209-250.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: Dual Route Cascaded Model of visual word recognition and reading aloud. *Psychological Review, 108*, 204-256.

Conway, A. R. A., Kane, M., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*(5), 769-786.

Corder, S. P. (1967). The significance of learners' errors. *IRAL, 5,* 161-170.

Coughlin, C. E., & Tremblay, A. (2013). Proficiency and working memory based explanations for nonnative speakers' sensitivity to agreement in sentence processing. *Applied Psycholinguistics*, *34*, 615–646.

Cowan, N. (2005). *Working memory capacity*. New York, NY: Psychology Press.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671-684.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104,* 268-294.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd Ed.). Thousand Oaks, CA: Sage Publications, Inc.

Croft, W. (1995). Modern syntactic typology. In M. Shibatani & T. Bynon (Eds.), *Approaches to language typology* (pp. 85-144). New York: Clarendon Press.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research, 28*(3), 369-382.

Cunnings, I., & Sturt, P. (2014). Coargumenthood and the processing of reflexives. *Journal of Memory and Language, 75*, 117-139.

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*, 450-466.

Daneman, M., & Green, I. (1986). Individual differences in comprehending and producing words in context. *Journal of Memory and Language, 25*, 1-18.

Davies, A. (1984). Simple, simplified and simplification: What is authentic? In J. Alderson, & A. Urquhart (Eds.), *Reading in a foreign language* (pp. 181-198). New York: Longman.

Davis, J. N. (1998). Facilitating effects of marginal glosses on foreign language reading. *Modern Language Journal, 73*(1), 41-48.

Davis, J. N., & Lyman Hager, M. (1997). Computers and L2 learning: Student performance, student attitudes. *Foreign Language Annals, 30*(1), 58-72.

de Bot, K., Lowie, W., & Verspoor, M. (2005). *Second language acquisition: An advanced resource book*. London, UK: Routledge.

de Bot, K., Lowie, W., & Verspoor, M. (2007). A dynamic systems approach to second language acquisition. *Bilingualism: Language and Cognition, 10*, 7-21, 51-55.

de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional Science, 38*, 105-134.

Dehaene. S., & Changeux, J. P. (2004). Neural mechanisms for access to consciousness. In M. Gazzaniga (Ed.), *The cognitive neurosciences III* (pp. 1145-1158). Cambridge, MA:MIT Press.

DeKeyser, R. (1998). Beyond focus on form: Cognitive perspectives on learning and practicing second language grammar. In C. Doughty, & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 42–63). Cambridge, UK: Cambridge University Press.

DeKeyser, R. M. (2005). What makes learning second-language grammar difficult? A review of issues. *Language Learning, 55*(Suppl. 1), 1-25.

De Zeeuw, M., Verhoeven, L., Schreuder, R. (2012). Morphological family size effects in young first and second language learners: Evidence of cross-language semantic activation in visual word recognition. *Language Learning, 62*(1), 68-92.

Dienes, Z., & Scott, R. (2005). Measuring unconscious knowledge: distinguishing structural knowledge and judgment knowledge. *Pyschological Research, 69*, 338-351.

Doddis, A. (1985). La cohesion textual en un discurso expositivo autentico y simplificado y en sus correspondientes recreaciones. *Lenguas Modernas, 12*, 136-148.

Dörnyei, Z. (2002). The motivational basis of language learning tasks. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 137-158). Amsterdam, The Netherlands: John Benjamins.

Dörnyei, Z. (2009). Individual differences: interplay of learner characteristics and learning environment. *Language Learning, 59*, 230-248.

Dörnyei, Z., & Kormos, J. (2000). The role of individual and social variables in oral task performance. *Language Teaching Research, 4*, 275-300.

Doughty, C. (1991). Second language instruction does make a difference: Evidence from an empirical study of SL relativization. *Studies in Second Language Acquisition, 13*, 431-469.

Doughty, C. (2001). Cognitive underpinnings of focus on form. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 206-257). New York, NY: Cambridge University Press.

Doughty, C., & Williams, J. (1998). *Focus on form in classroom second language acquisition*. Cambridge, UK: Cambridge University Press.

Dulay, H. C., & Burt, M. K. (1973). Should we teach children syntax? *Language Learning, 23*, 245-258.

Dussias, P. E., Valdés Kroff, J. R., Tamargo, R. E. G., Gerfen, C. (2013). When gender and looking go hand in hand: Grammatical gender processing in L2 Spanish. *Studies in Second Language Acquisition, 35,* 353-387.

Educational Testing Services (2013). *Official TOEFL iBT Tests with Audio Volume 1*. McGraw-Hill: New York.

Egi, T. (2004). Verbal reports, noticing, and SLA research. *Language Awareness, 13*, 243-264.

Egi, T. (2008). Investigating stimulated recall as a cognitive measure: Reactivity and verbal reports in SLA research methodology. *Language Awareness, 17*(3), 212-217.

Egi, T., Adams, R. J., & Nuevo, A. (2013). Is metalinguistic stimulated recall reactive in second language learning? In J. M. Bergsleithner, S. N. Frota, & J. K. Yoshioka (Eds.), *Noticing and second language acquisition: Studies in Honor of Richard Schmidt* (pp. 81-102). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition, 18*, 91-126.

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition, 27*, 305-352.

Ellis, N. C. (2012) Frequency-based accounts of SLA. In S. Gass & A. Mackey (Eds.), *Handbook of second language acquisition* (pp. 193-210). New York, NY:Routledge.

Ellis, N. C., & Sinclair, S. G. (1996). Working memory in the acquisition of vocabulary and syntax: Putting language in good order. *The Quarterly of Experimental Psychology, 49A*(1), 234-250.

Ellis, N. C., & Schmidt, R. (1997). Morphology and longer distance dependencies: Laboratory research illuminating the A in SLA. *Studies in Second Language Acquisition, 19*, 145-171.

Ellis, R. (2003). *Task-based language learning and teaching.* Oxford, UK: Oxford University Press.

Ellis, R. (2009). Task-based language teaching: Sorting out the misunderstandings. *International Journal of Applied Linguistics, 19*(3), 221–246.

Ellis, R., Loewen, S., Elder, C., Erlam, R., Philp, J., & Reinders, H. (2009). *Implicit and explicit knowledge in second language learning, testing and teaching*. Bristol: Multilingual Matters.

Engel de Abreu, P. M. J., & Gathercole, S. E. (2012). Executive and phonological processes in second-language acquisition. *Journal of Educational Psychology*, *104*, 974–986.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal report as data*. Cambridge, MA: MIT.

Eskey, D. E. (2005). Reading in a second language. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 563-579). Mahwah, NJ: Lawrence Erlbaum.

Fender, M. (2001). A review of L1 and L2/ESL word integration skills and the nature of L2/ESL word integration development involved in lower-level text processing. *Language learning, 51*, 319-396.

Ferguson, C. (2009). An effect size primer: A guide to clinicians and researchers. *Profes- sional Psychology: Research and Practice*, *40*, 532–538.

Field, J. (2004). *Psycholinguistics: The key concepts*. New York, NJ: L. Erlbaum.

Fink, A., & Neubauer, A. C. (2001). Speed of information processing, psychometric intelligence: and time estimation as an index of cognitive load. *Personality and Individual Differences, 30*, 1009-1021.

Foote, R. (2011). Integrated knowledge of agreement in early and late English-Spanish bilinguals. *Applied Psycholinguistics*, *32*, 187–220.

Foster, P., & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language Learning, 59*(4), 866-896.

Fox, J. (2002). *An R and S-Plus companion to applied regression*. Thousand Oaks, CA: Sage.

Fredericks, T.K., Choi S.D., Hart J., Butt S.E., & Mital A. (2005). An investigation of myocardial aerobic capacity as a measure of both physical and cognitive workloads. *International Journal of Industrial Ergonomics, 35*(12), 1097–1107.

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing, 10*, 133-170.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing, 16*(1), 2-32.

French, L. M. (2006). *Phonological working memory and second language acquisition: A developmental study of Francophone children learning English in Quebec*. Lewiston, NY: Edwin Mellen Press.

French, L. M., & O'Brien, I. (2008). Phonological memory and children's second language grammar learning. *Applied Psycholinguistics*, *29*, 463–487.

Frenck-Mestre, C. (2005). Eye-movement as a tool for studying syntactic processing in a second language: A review of methodologies and experimental findings. *Second Language Research, 21*, 175–198.

Gagné, C.L., & Spalding, T.L. (2009). Constituent integration during the processing of compound words: Does it involve the use of relational structures? *Journal of Memory and Language*, *60*, 20-35.

Gao, X., & Gu, X. (2008). An introspective study on test-taking process of banked cloze. *CELEA Journal*, *31*(4), 3–16.

García Mayo, M. P. & Azkarai, A. (2016). EFL task-based interaction: Does task modality impact on language-related episodes? In M. Sato & S. Ballinger (Eds.), *Peer interaction and second language learning pedagogical potential and research agenda* (pp. 241–266). Amsterdam, The Netherlands: John Benjamins.

Gass, S. M. (1997). *Input, interaction, and the second language learner*. Mahwah, NJ: Lawrence Erlbaum.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah: NJ: Erlbaum.

Gathercole, S. E. (1995). Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. *Memory & Cognition, 23*(1), 83-94.

Gathercole, S. E., Pickering, S. J., Hall, M., & Peaker, S. M. (2001). Dissociable lexical and phonological influences on serial recognition and serial recall. *The Quarterly Journal of Experimental Psychology, 54A*(1), 1-30.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gettys, S., Imhof, L., & Kautz, J. O. (2001). Computer-assisted reading: The effect of glossing format on comprehension and vocabulary retention. *Foreign Language Annals, 34*(2), 91-99.

Geva, E., & Ryan, E. B. (1993). Linguistic and cognitive correlates of academic skills in first and second language. *Language Learning, 43*, 5-42.

Gilabert, R. (2007). Effects of manipulating task complexity on self-repairs during L2 oral production. *IRAL, 45*, 215-240.

Gilabert, R., Barón, J., & Llanes, À. (2009). Manipulating cognitive complexity across task types and its impact on learners' interaction during oral performance. *IRAL, 47*, 367-395.

Gilabert, R., Barón, J., & Levkina, M. (2011). Manipulating task complexity across task types and modes. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (pp. 105-138). Amsterdam, The Netherlands: John Benjamins.

Gilabert, R., Manchón, R., & Vasylets, O. (2016). Mode in theoretical and empirical TBLT research: Advancing research agendas. *Annual Review of Applied linguistics, 36*, 117-135.

Godfroid, A., & Uggen, M. S. (2013). Attention to irregular verbs by beginning learners of German: an eye-movement study. *Studies in Second Language Acquisition, 35*, 291-322.

Godfroid, A., Housen, A., & Boers, F. (2010). A procedure for testing the Noticing Hypothesis in the context of vocabulary acquisition. In M. Pütz & L. Sicola (Eds.), *Inside the learner's mind: Cognitive processing and second language acquisition* (pp. 169-197). Amsterdam/Philadelphia, PA: John Benjamins.

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye-tracking. *Studies in Second Language Acquisition, 35*, 483-517.

Godfroid, A., Winke, P., & Gass, S. (2013). Special issue: Eye-movement recordings in second language research. *Studies in Second Language Acquisition, 35*, 205-422.

Godfroid, A., Loewen, S., Jung, S., Park, J., & Gass, S. (2015). Timed and untimed grammaticality judgments measure distinct types of knowledge: Evidence from eye-movement patterns. *Studies in Second Language Acquisition, 37*, 269-297.

Goldberg, A. E. (1998). Patterns of experience in patterns of language. In M. Tomasello (Ed.), *The new psychology of language: Cognitive and functional approaches to language structure* (pp. 203-219). Mahwah, NJ: Erlbaum.

Goo, J. (2010). Working memory and reactivity. *Language Learning, 60*(4), 712-752.

Goo, J. (2012). Corrective feedback and working memory capacity in interaction-driven L2 learning. *Studies in Second Language Acquisition, 34*, 445-474.

Gorsuch, G., & Taguchi, E. (2010). Developing reading fluency and comprehension using repeated reading: Evidence from longitudinal student reports. *Language Teaching Research, 14*(1), 27-59.

Grabe, W. (2005). The role of grammar in reading comprehension. In J. Frodesen & C. Holton (Eds.), *The power of context in language teaching and learning* (pp. 268-282). Boston: Heinle & Heinle.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge, UK: Cambridge University Press.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge, UK: Cambridge University Press.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychologics*. New York, NY: Wiely.

Gries, S.T. (2013). *Statistics for linguistics with R: A practical introduction*. Berlin: De Gruyter.

Guidi, C. (2009). *Glossing for meaning and glossing for form. A computerized study of the effects of glossing and type of linguistic item on reading comprehension,*

noticing, and L2 learning. Unpublished dissertation, Georgetown University, Washington, D.C.

Gurzynski-Weiss, L., & Baralt, M. (2014). Exploring learner perception and use of task-based interactional feedback in FTF and CMC modes. *Studies in Second Language Acquisition, 36*, 1-37.

Halliday, M., & Matthiessen, C. (2013). *An introduction to functional grammar.* London: Hodder Arnold.

Han, Z., & Cadierno, T. (2010). *Linguistic relativity in SLA: Thinking for speaking.* Bristol: Multilingual Matters.

Han, Z., Park, E. S., & Combs, C. (2008). Textual enhancement of input: Issues and possibilities. *Applied Linguistics, 29*(4), 597-618.

Han, Z., Anderson, N., & Freeman, D. (2009). Introduction: Crossing the boundaries. In Z. Han & N. Anderson (Eds.), *Second language reading: Research and instruction* (pp. 1-13). Ann Arbor, MI: University of Michigan Press.

Harrell, F. E., & Dupont, C. (2015). *Hmisc: Harrell Miscellaneous*. URL https://CRAN.R-project. org/package=Hmisc. R package version 3.17-0.

Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. *Studies in Second Language Acquisition, 14*, 25-38.

Hatch, E. (1983). Simplified input and second language acquisition. In R. W. Anderson (Ed.), *Pidginization and creolization as language acquisition* (pp. 64-86). Rowley, MA: Newbury House.

Heift, T., & Rimrott, A. (2012). Task-related variation in computer-assisted language learning. *Modern Language Journal, 96*(4), 525-543.

Henderson, J. M., & Ferreira, E. (1990). Effects of foveal processing difficulty on the perceptual span in reading: Implications for attention and eye movement

control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 417–429.

Hicks, R. E., Miller, G. W., & Kinsbourne, M. (1976). Prospective and retrospective judgments of time as a function of amount of information processed. *American Journal of Psychology, 89*(4), 719-730.

Hoover, M. L., & Dwivedi, V. D. (1998). Syntactic processing by skilled bilinguals. *Language Learning, 48*, 1-29.

Horiba, Y. (2000). Reader control in reading: Effects of language competence, text type, and task. *Discourse Processes, 29*, 223-267.

Horiba, Y. (2013). Task-induced strategic processing in L2 text comprehension. *Reading in a Foreign Language, 25*(2), 98-125.

Huang, L., & Lin, C. (2014). Three approaches to glossing and their effects on vocabulary learning. *System, 44,* 127-136.

Hulstijn, J. H. (1992). Retention of inferred and given word meanings: Experiment in incidental vocabulary learning. In. P. J. L. A. a. H. Bejoint (Ed.), *Vocabulary and applied linguistics* (pp. 113-125). London, UK: McMillan.

Hulstijn, J. H. (1995). Not all grammar rules are equal: Giving grammar instruction its proper place in foreign language teaching. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 359-386). Honolulu, HI: University of Hawaii Press.

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning, 51*(3), 539-558.

Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *Modern Language* Journal, 80(3), 327-339.

Hummel, K. M. (2009). Aptitude, phonological memory, and second language proficiency in non-novice adult learners. *Applied Psycholinguistics*, *30*, 225–249.

Hwang, J. B. (1999). L2 acquisition of English unaccusative verbs under implicit and explicit learning conditions. *English Teaching, 54*(4), 145-176.

Hwang, J. B. (2001). Focus on form and the L2 learning of English unaccusative verbs. *English Teaching, 56*(3), 111-133.

Indrarathne, H. D. B. N., & Kormos, J. (in press). Relationship between attentional processing of input and working memory: an eye-tracking study. *Studies in Second Language Acquisition.*

Ishikawa, T. (2007). The effect of increasing task complexity along the [±Here-and-Now] dimension on L2 written narrative discourse. In M. P. Garcia Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 136-156). Clevedon, UK: Multilingual Matters.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning, 51*(3), 401-436.

Izumi, S. (2002). Output, input enhancement, and the noticing hypothesis: An experimental study on ESL relativization. *Studies in Second Language Acquisition, 24*, 541-577.

Izumi, S, (2003). Comprehension and production processes in second language learning: In search of the psycholinguistic rationale of the Output Hypothesis. *Applied Linguistics, 24*(2), 168-196.

Izumi, S., Bigelow, M., Fujiwara, M., & Fearnow, S. (1999). Testing the Output Hypothesis: Effects of output on noticing and second language acquisition. *Studies in Second Language Acquisition, 21*, 421-452.

Jackson, D. O., & Suethanapornkul, S. (2013). The cognition hypothesis: A synthesis and meta-analysis of research on second language task complexity. *Language Learning, 63*(2), 330-367.

Jackson, C. N., Dussias, P. E., & Hristova, A. (2012). Using eye-tracking to study the on-line processing of case-marking information among intermediate L2 learners of German. *International Review of Applied Linguistics, 50*(2), 101-133.

Jacobs, G., Dufon, P., & Hong, F. C. (1994). L1 and L2 glosses in reading passages, their effectiveness for increasing comprehension and vocabulary knowledge. *Journal of Research in Reading 17*(1), 19-28.

Jahan , A., & Kormos, J. (2015). The impact of textual enhancement on EFL learners' grammatical awareness of future plans and intentions. *International Journal of Applied Linguistics, 25*, 46-66.

Jarrold, C., Baddeley, A. D., & Hewes, A. K. (1999). Dissociating working memory: Evidence from Down's and Williams syndrome. *Neuropsychologia, 37*, 637-651.

Jarrold, C., & Towse, J. N. (2006). Individual differences in working memory. *Neuroscience*, *139*, 39–50.

Jarvis, S., & Pavlenko, A. (2008). *Cross-linguistic influence in language and cognition*. London, UK: Routledge.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. New York, NY: Springer-Verlag.

Jiang, N. (2011). *Conducting reaction time research in second language studies*. New York, NY: Routledge.

Johnson, P. (1982). Effects on reading comprehension of building background knowledge. *TESOL Quarterly, 16*(4), 503-516.

Jourdenais, R. (2001). Protocol analysis and SLA. In P. Robinson (Ed.), *Cognition and second language acquisition* (pp.354-375). New York: Cambridge.

Jourdenais, R., Ota, M., Stauffer, S., Boysen, B., & Doughty, C. (1995). Does textual enhancement promote noticing? A think aloud protocol analysis. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 183-216). Honolulu: University of Hawai'i, National Foreign Language Resource Center.

Ju, M. K. (2000). Overpassivization errors by second language learners: The effect of conceptualizable agents in discourse. *Studies in Second Language Acquisition, 22*, 85-111.

Juffs, A. (2001). Psycholinguistically oriented second language research. *Annual Review of Applied Linguistics, 21,* 207-221.

Juffs, A. (2004). Representation, processing and working memory in a second language. *Transactions of the Philological Society, 102*, 199-226.

Juffs, A. (2005). The influence of first language on the processing of *wh*-movement in English as a second language. *Second Language Research, 21*, 121-151.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 98, 122–149.

Kahneman, D., & Beatty, J. (1967). Pupillary responses in a pitch-discrimination task. *Perception and Psychophysics, 2*, 101-105.

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology, 13*, 351-371.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O, Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133*(2), 189-217.

Kaushanskaya, M., & Marian, V. (2007). Bilingual language processing and interference in bilingualism: Evidence from eye tracking and picture naming. *Language Learning, 57*(1), 119-163.

Keating, G. D. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research, 12,* 365-386.

Keating, G. D. (2009). Sensitivity to violations of gender agreement in native and nonnative Spanish: An eye-movement investigation. *Language Learning, 59*(3), 503-353.

Kempe, V., & Brooks, P. J. (2008). Second language learning of complex inflectional systems. *Language Learning, 58*(4), 703-746.

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language learning.* Cambridge, UK: Cambridge University Press.

Kieffer, M. J., & Lesaux, N. K. (2012). Direct and indirect roles of morphological awareness in the English reading comprehension of native English, Spanish, Filipino, and Vietnamese speakers. *Language Learning, 62*(4), 1170-1204.

Kim, J., & Kim, H. S. (2012). Korean L2 overpassivization of unaccusative verbs: Focusing on the relative effects of linguistic and cognitive factors. *The Journal of English Language and Literature, 54*(4), 155-183.

Kim, Y-J. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 58*(2), 285-325.

Kim, Y-J. (2009). The effects of task complexity on learner-learner interaction. *System, 37*, 254-268.

Kim, Y-J. (2012). Task complexity, learning opportunities and Korean EFL learners' question development. *Studies in Second Language Acquisition, 34*, 627-658.

Kim, Y-J., & Tracy-Ventura, N. Task complexity, language anxiety, and the development of the simple past. (2011) In P. Robinson (Ed.), *Researching second language task complexity: Task demands, language learning and language performance* (pp. 287-306). Amsterdam, The Netherlands: John Benjamins.

Kim, Y-J., & Taguchi, N. (2015). Promoting task-based pragmatics instruction in EFL classroom contexts: The role of task complexity. *Modern Language Journal, 99*, 4, 656-677.

Kim,Y-J., Payant, C., & Pearson, P. (2015). The intersection of task-based interaction, task complexity, and working memory: L2 question development through recasts in a laboratory setting. *Studies in Second Language Acquisition, 37*, 549-581.

Kintsch, W. (1998). *Comprehension: A framework for cognition*. New York: Cambridge University Press.

Kintsch, W., & van Dijk. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85, 363-394.

Ko, H. M. (2005). Glosses, comprehension, and strategy use. *Reading in a Foreign Language, 17*(2), 125-143.

Ko, H. M. (2012). Glossing and second language vocabulary learning. *TESOL Quarterly, 46*(1), 56-79.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: text organization and response format. *Language Testing, 19*(2), 193-220.

Kormos, J. (2011). Speech production and the cognition hypothesis. In P. Robinson (Ed.), *Researching second language task complexity: Task demands, language learning and language performance* (pp. 39-60). Amsterdam, The Netherlands: John Benjamins.

Kormos J., & Dörnyei, Z. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*, 146-164.

Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition, 11*(2), 261-271.

Kormos, J., & Trebits, A. (2011). Working memory capacity and narrative task performance. In P. Robinson (Ed.), *Researching second language task complexity: Task demands, language learning and language performance* (pp. 267-285). Amsterdam, The Netherlands: John Benjamins.

Kormos, J., & Trebits, A. (2012). The role of task complexity, modality, and aptitude in narrative task performance. *Language Learning, 62*(2), 439-472.

Kost, C. R., Foss, P. & Lenzini, J. J. (1999). Textual and pictorial glosses: Effectiveness on incidental vocabulary growth when reading in a foreign language. *Foreign Language Annals, 32*(1), 89-113.

Krashen, S. (1982). *The input hypothesis*. Oxford, UK: Pergamon Press.

Kuiken, F., & Vedder, I. (2005). Cognitive task complexity and second language writing performance. In S. Foster-Cohen, M. P. García Mayo, & J. Cenoz (Eds.), *Eurosla Yearbook* (Vol. 5, pp. 195-222). Amsterdam, The Netherlands: John Benjamins.

Kuiken, F., & Vedder, I. (2007a). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing, 17*, 48-60.

Kuiken, F., & Vedder, I. (2007b). Task complexity and measures of linguistic performance in L2 writing. *International Review of Applied Limguistics, 45*, 261-284.

Kuiken, F., & Vedder, I. (2011). Task complexity and linguistic performance in L2 writing and speaking. In P. Robinson (Ed.), *Researching second language task complexity: Task demands, language learning and language performance* (pp. 91-104). Amsterdam, The Netherlands: John Benjamins.

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition, 10*, 294-340.

LaBrozzi, R. (2016). The effects of textual enhancement type on L2 form recognition and reading comprehension in Spanish. *Language Teaching Research, 20*(1), 75-91.

Lahtinen, T. M., Koskelo, J. P., Laitinen, T., & Leino, T. K. (2007). Heart rate and performance during combat missions in a flight simulator. *Aviation, Space, and Environmental Medicine, 78*(4), 387-395.

Lai, C., Zhao, Y., & Wang, J. (2011). Task-based language teaching in online Ab Initio foreign language classrooms. *The Modern Language Journal, 95*, 81-103.

Larsen-Freeman, D. (2010). The dynamic co-adaptation of cognitive and social views: A complexity theory perspective. In R. Batstone (Ed.), *Sociocognitive perspective on language use and language learning* (pp. 40-53). Oxford, UK: Oxford University Press.

Larsen-Freeman, D. (2012). Complexity theory. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 73-87). London, UK: Routledge.

Laufer, B. (2009). Second language vocabulary acquisition from language input and from form focused activities. *Language Teaching, 42*, 341-354.

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics, 22*, 1-26

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54*(3), 399-436.

Lee, S.-K. (2007). Effects of textual enhancement and topic familiarity on Korean EFL students' reading comprehension and learning of passive form. *Language Learning, 57*, 87-118.

Lee, S.-K., Miyata, M., & Ortega, L. (2008). A usage-based approach to overpassivization: The role of input and conceptualization biases. *Paper presented at the 26th Second Language Research Forum*, Honolulu, HI, October 17-19.

Leeman, J., Arteagoitia, I., Fridman, B., & Doughty, C. (1995). Integrating attention to form with meaning: Focus on form in content-based Spanish instruction. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 217-258). Honolulu: University of Hawai'i, National Foreign Language Resource Center.

Leeser, M. J. (2007). Learner-based factors in L2 reading comprehension and processing grammatical form: topic familiarity and working memory. *Language Learning, 57*(2), 229-270.

Leow, R. P. (1993). To simplify or not to simplify. *Studies in Second Language Acquisition, 15*, 333-355.

Leow, R. P. (1997a). Attention, awareness and foreign language behavior. *Language Learning, 47*, 467-505.

Leow, R. P. (1997b). The effects of input enhancement and text length on adult L2 readers' comprehension and intake in second language acquisition. *Applied Language Learning, 8,* 151-182.

Leow, R. P. (1997c). Simplification and second language acquisition. *World Englishes, 16*, 291-296.

Leow, R. P. (2000). A study of the role of awareness in foreign language behavior: Aware versus unaware learners. *Studies in Second Language Acquisition, 22*(4), 557-584.

Leow, R. P. (2001a). Attention, awareness, and foreign language behavior. *Language Learning, 51*(Suppl 1), 113-155.

Leow, R. P. (2001b). Do learners notice enhanced form while interacting with the L2? An online and offline study of the role of written input enhancement in L2 reading. *Hispania, 84*, 496-509.

Leow, R. P. (2009). Modifying the L2 reading text for improved comprehension and acquisition: Does it work? In Z.-H. Han & N. J. Anderson (Eds.), *Second language reading research and instruction: Crossing the boundaries* (pp. 83-100). Ann Arbor, MI: The University of Michigan Press.

Leow, R. P. (2015). *Explicit learning in the L2 classroom: A student-centered approach*. New York, NY: Routledge.

Leow, R. P., & Hama, M. (2013). Implicit learning in SLA and the issue of internal validity: A response to Leung and Williams's (2011) "The implicit learning of mappings between forms and contextually derived meanings." *Studies in Second Language Acquisition, 35*, 545-557.

Leow, R. P., & Morgan-Short, K. (2004). To think aloud or not to think aloud: The issue of reactivity in SLA research methodology. *Studies in Second Language Acquisition, 26,* 35-57.

Leow, R. P., Egi, T., Nuevo, A. M., & Tsai, Y. (2003). The roles of textual enhancement and type of linguistic item in adult L2 learners' comprehension and intake. *Applied Language Learning, 13(*2), 1-16.

Leow, R. P., Hsieh, H., & Moreno, N. (2008). Attention to form and meaning revisited. *Language Learning, 58*(3), 665-695.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levelt, W. J. M. (1999). Language production: a blueprint of the speaker. In C. Brown & P. Hagoort (Eds.), *Neurocognition of language* (pp. 83-122). Oxford: Oxford University Press.

Li, S. (2013). The interactions between the effects of implicit and explicit feedback and individual differences in language analytic ability and working memory. *Modern Language Journal, 97(*3), 634-654.

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning, 65*(1), 185-207.

Loewen, S., & Inceoglu, S. (2016). The effectiveness of visual input enhancement on the noticing and L2 development of the Spanish past tense. *Studies in Second Language Learning and Teaching,* (VI-1), 89-110.

Lomicka, L. L. (1998). "To gloss or not to gloss": An investigation of reading comprehension online. *Language learning & Technology, 1*(2), 41-50.

Long, M. H. (1985). Input and second language acquisition theory. In S. Gass & C. Madden (Eds.), *Input in second language acquisition*. (pp. 377-393) Rowley, MA: Newburry House.

Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. In K. de Bot, R. Ginsberg & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspective* (pp. 39–52). Amsterdam, The Netherlands: Benjamins.

Long, M. H. (2016). In defense of tasks and TLBT: Non-issues and real issues. *Annual Review of Applied Linguistics, 36*, 5-33.

Long, M. H., & Crookes, G. (1992). Three approaches to task-based language teaching. *TESOL Quarterly* 26, 1, 27-56.

Long, M. H., & Ross, S. (1997). Modifications that preserve language and content. In M. L. Tickoo (Ed.), *Simplification: Theory and application* (pp. 29-52). Singapore: SEAMEO Regional Language Centre.

Loschky, L., & Bley-Vroman, R. (1993). Grammar and task-based methodology. In G. Crookes & S. Gass (Eds.), *Tasks and language learning: Integrating theory and practice* (pp. 123-167). Clevedon, England: Multilingual Matters Ltd.

Mackey, A. (2006). Feedback, noticing and instructed second language learning. *Applied Linguistics, 27*(3), 405-430.

Mackey, A., & Sachs, R. (2012). Older learners in SLA research: A first look at working memory, feedback, and L2 development. *Language Learning, 62*(3), 704-740.

Mackey, A., Gass, S. M., & McDonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition, 22*, 471-497.

Mackey, A., Philp, J., Egi, T., Fujii, A., & Tatsumi, T. (2002). Individual differences in working memory, noticing of interactional feedback and L2 development. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 181-209). Amsterdam, The Netherlands: John Benjamins.

Mackey, A., Adams, R., Stafford, C., & Winke, P. (2010). Exploring the relationship between modified output and working memory capacity. *Language Learning, 60*(3), 501-533.

MacWinney, B. (2012). The logic of the unified model. In S. Gass & A. Mackey (Eds.), *Handbook of second language acquisition* (pp. 211-227). New York, NY: Routledge.

Martin, K. I., & Ellis, N. C. (2012). The roles of phonological short-term memory and working memory in L2 grammar and vocabulary learning. *Studies in Second Language Acquisition, 34*(3), 379-413.

Martinez-Fernández, A. M. (2010). *Experiences of remembering and knowing in SLA, L2 development, and text comprehension: A study of levels of awareness, type of glossing, and type of linguistic item*. Unpublished dissertation, Georgetown University, Washington, D.C.

Masoura, V. M., & Gathercole, S. E. (2005). Phonological short-term memory skills and new word learning in young Greek children. *Memory, 13*, 422-429.

McCray, G. (personal communication, August 9, 2016). http://rpubs.com/GarethMcCray/reading-metrics.

McLaughlin, B. (1987). *Theories of second language learning*. London, UK: Edward Arnold.

Meyer, B. (1975). *The organization of prose and its effects on memory*. Amsterdam, The Netherlands: North-Holland.

Meyer, B. (1985). Prose analysis: Purposes, procedures, and problems. In B. Britton & J. Black (Eds.), *Analyzing and understanding expository text* (pp. 11-64, 269-304). Hillsdale, NJ: L. Erlbaum.

Meyer, B., & Freedle, R. O. (1984). Effects of discourse type on recall. *American Educational Research Journal, 21,* 121-143.

Meyer, B. J. F., & Ray, M. N. (2011). Structure strategy interventions: Increasing reading comprehension of expository text. *International Electronic Journal of Elementary Education, 4*(1), 127-152.

Michel, M. C. (2011). Effects of task complexity and interaction on L2 performance. In P. Robinson (Ed.), *Researching second language task complexity: Task demands, language learning and language performance* (pp. 141-173). Amsterdam, The Netherlands: John Benjamins.

Michel, M. C. (2013). The use of conjunctions in cognitively simple versus complex oral L2 tasks. *Modern Language Journal, 97*(1), 178-195.

Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics, 45*, 241-259.

Miyake, A. (2001). Individual differences in working memory: Introduction to the special section. *Journal of Experimental Psychology: General*, *130*, 163–168.

Miyake, A., & Friedman, D. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. F. Healy & L. E. Bourne, Jr. (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 339-364). Mahwah, NJ: Erlbaum.

Morgan-Short, K., Heil, J., Botero-Moriarty, A., & Ebert, S. (2012). Allocation of attention to second language form and meaning: Issues of think-alouds and depth of processing. *Studies in Second Language Acquisition, 34*, 659-685.

Murata, A. (2005). An attempt to evaluate mental workload using wavelet transform of EEG. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 47*, 498-508.

Nagata, N., (1999). The effectiveness of computer-assisted interactive glosses. *Foreign Language Annals, 32*, 4, 469-479

Nassaji, H. (2003). Higher-level and lower-level processing skills in advanced ESL reading comprehension. *Modern Language Journal*, *87*, 261-276.

Nassaji, H. (2007). Schema theory and knowledge-based processes in second language reading comprehension: A need for alternative perspectives. *Language Learning, 57*, 79-113.

Nassaji, H. (2014). The role and importance of lower-level processes in second language reading. *Language Teaching, 47*(1), 1-37.

Nieuwenhuis, R., Pelzer, B., & te Grotenhuis, M. (2012). *Influence. ME: Tools for detecting influential data in mixed effects models.* URL http://CRAN. R-porject. Org/package=influence.ME.

No, G., & Chung, T. (2006). Multiple effects and the learnability of English unaccusatives. *English Teaching, 61*(1), 19-39.

Norris, J. M. (1996). *A validation study of the ACTFL guidelines and the German speaking test*. Unpublished MA thesis, University of Hawai'i.

Norris, J. M. (2009). Task-based teaching and testing. In M. H. Long & C. J. Doughty (Eds.), *Handbook of language teaching* (pp. 578–594). Oxford, UK: Blackwell.

Norris, J. M., & Ortega, L. (2003). Defining and measuring SLA. In C. J. Doughty & M. H. Long, (Eds.), *Handbook of second language acquisition* (pp. 716-761). London, UK: Blackwell.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555-578.

Norris, J. M., Bygate, M. & van den Branden, K. (2009). Task-based language assessment. In K. van den Branden, M. Bygate, & J.M. Norris (Eds.), *Task-based language teaching. A reader* (pp. 431–434). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Nuevo, A. (2006). *Task complexity and interaction: L2 learning opportunities and interaction*. Unpublished doctoral dissertation. Georgetown University, Washington D.C.

Nuevo, A-M., Adams, R., & Ross-Feldman, L. (2011). Task complexity, modified output, and L2 development in learner-learner interaction. In P. Robinson (Ed.), *Researching second language task complexity: Task demands, language*

learning and language performance (pp. 175-201). Amsterdam, The Netherlands: John Benjamins.

Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge, UK: Cambridge University Press.

Nunan, D. (2004). *Task-based language teaching*. Cambridge, UK: Cambridge University Press.

O'Brien, I., Segalowitz, N., Collentine, J., & Freed, B. (2006). Phonological memory and lexical, narrative, and grammatical skills in second language oral production by adult learners. *Applied Psycholinguistics, 27*, 377-402.

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning*. Cambridge, UK: Cambridge University Press.

Oh, S-Y. (2001). Two types of input modification and EFL reading comprehension: Simplification versus elaboration. *TESOL Quarterly, 35*(1), 69-96.

Ono, K., & Budwig, N. (2006). Young children's use of unaccusative intransitives in novel verb experiments. In D. Bamman, T. Magnitskaia, & C. Zaller (Eds.), *A supplement to the proceedings of the 30th Boston University Conference on Language*. (Retrievable from: http://www.bu.edu/linguistics/APPLIED/BUCLD/supp30.html)

Osaka, M., & Osaka, N. (1992). Language-independent working memory as measured by Japanese and English reading span tests. *Bulletin of the Psychonomic Society, 30*, 287-289.

Osaka, M., Osaka, N., & Groner. R. (1993). Language-independent working memory: Evidence from German and French reading span tests. *Bulletin of the Psychonomic Society, 31*, 117-118.

Oshita, H. (2000). What is happened may not be what appears to be happening: A corpus study of 'passive' unaccusatives in L2 English. *Second Language Research, 16*(4), 293-324.

Overstreet, M. H. (1998). Text enhancement and content familiarity: The focus of learner attention. *Spanish Applied Linguistics, 2*, 229-258.Paas, F. G. W. C., & van Merriënboer, J. J. G. (1994a). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review, 6*(4), 351-371.

Paas, F. G. W. C., & van Merriënboer, J. J. G. (1994b). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86*(1), 122-133.

Pass, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63-71.

Pak. J. (1986). *The effect of vocabulary glossing on ESL reading comprehension*. Unpublished manuscript, University of Hawaii at Manoa.

Park, E. S. (2004). Constraints of implicit focus on form: insights from a study of input enhancement. *Teachers College, Columbia University, Working Papers in TESOL & Applied linguistics, 4*(2), http://journals.tc-library.org/index.php/tesol/articles/view File/59/65.

Park, E. S., & Nassif, L. (2014). Textual enhancement of two L2 Arabic forms: A classroom-based study. *Language Awareness, 23*(4), 334-352.

Pellicer-Sánchez. A. (2016). Incidental L2 vocabulary acquisition *from* and *while* reading. *Studies in Second Language Acquisition, 38*, 97-130.

Perfetti, C. A. (1999). Comprehending written language: A blueprint of the reader. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 167-210). Oxford: Oxford University Press.

Perlmutter, D. M. (1978). Impersonal and the unaccusative hypothesis. In *Proceedings of the 4th Annual Meeting of the Berkeley Linguistics Society* (pp. 157-190). Berkeley, CA: University of California, Berkeley Linguistics Society.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language testing*, *20*(1), 26–56.

Philp, J., & Iwashita, N. (2013). Talking, turning in and noticing: Exploring the benefits of output in task-based peer interaction. *Language Awareness, 22*(4), 353-370.

Pica, T. (1987). Second-language acquisition, social interaction, and the classroom. *Applied Linguistics, 8*, 3-21.

Pica, T. (2002). Subject-matter content: How does it assist the interactional and linguistic needs of classroom language learners? *Modern Language Journal, 86*, 1-19.

Pienemann, M. (1989). Is language teachable? Psycholinguistic experiments and hypotheses. *Applied Linguistics, 10*(1), 52-79.

Plonsky, L., & Kim, Y. (2016). Task-based learner production: A substantive and methodological review. *Annual Review of Applied Linguistics, 36*, 73-97.

Plonsky, L., & Oswald, F. L. (2015). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878-912.

Posner, M. (1988). Structures and functions of selective attention. In. T. Boll & B. Bryant (Eds.), *Master lectures in clinical neuropsychology* (pp. 173-202). Washington DC: American Psychology Association.

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience, 13*, 25-42.

Powell, M. J. D. (2009). *The BOBYQA algorithm for bound constrained optimization without derivative*s. Department of Applied Mathematics and Theoretical Physics. Cambridge University.

Pressley, M. (2006). *Reading instruction that works*. New York: Guilford Press.

Pulido, D. (2007). The effects of topic familiarity and passage sight vocabulary on L2 lexical inferencing and retention through reading. *Applied Linguistics, 28*(1), 66-86.

Pulido, D. (2009). How involved are American L2 learners of Spanish in lexical input processing tasks during reading. *Studies in Second Language Acquisition, 31*, 31-58.

R Development Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Racsmány, M., Lukács, Á, & Pléh, Cs. (2005). A verbális munkamemória magyar nyelvû vizsgálóeljárásai [Verbal working memory testing procedures in Hungarian]. *Pszicholo´giai Szemle*, *60*, 479–506.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *The Quarterly Journal of Experimental Psychology, 62*, 1457-1506.

Rayner, K., & Pollatsek, A. (1989). The psychology of reading. Hillsdale, NJ: Erlbaum.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6*, 855-863.

Rebuschat, P. (2013). Methodological review article: Measuring implicit and explicit knowledge in second language research. *Language Learning, 63*(3), 595-626.

Rebushcat, P., & Williams, J. N. (2012). Implicit and explicit knowledge in second language acquisition. *Applied Psycholinguistics, 33*(4), 829-856.

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment, 28*(3), 164-171.

Révész, A. (2007). *Focus on form in task-based language teaching: Recasts, task complexity, and L2 learning*. Unpublished doctoral dissertation. Teachers College, Columbia University, New York.

Révész, A. (2009). Task complexity, focus on form, and second language development. *Studies in Second Language Acquisition, 31*, 437-470.

Révész, A. (2011). Task complexity, focus on L2 constructions, and individual differences: A classroom-based study. *Modern Language Journal, 95*, 162-181.

Révész, A. (2012). Working memory and the observed effectiveness of recasts on different L2 outcome measures. *Language Learning, 62*(1), 93-132.

Révész, A. (2014). Towards a fuller assessment of cognitive models of task-based learning: Investigating task-generated cognitive demands and processes. *Applied Linguistics, 35*(1), 87-92.

Révész, A., & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition, 35*, 31-65.

Révész, A., Sachs, R., & Mackey, A. (2011). Task complexity, uptake of recasts, and second language development. In P. Robinson (Ed.), *Researching second language task complexity: Task demands, language learning and language performance* (pp. 203-236). Amsterdam, The Netherlands: John Benjamins.

Révész, A., Sachs, R., & Hama, M. (2014). The effects of task complexity and input frequency on the acquisition of the past counterfactual construction through recasts. *Language Learning, 64*(3), 615-650.

Révész, A., Michel, M., & Gilabert, R. (2016). Measuring cognitive task demands using dual-task methodology, subjective self-ratings, and expert judgments: A validation study. *Studies in Second Language Acquisition*. Advance online publication. doi:10.1017/S0272263115000339.

Révész, A., Kourtali, N-E., Mazgutova, D. (in press). Effects of task complexity on L2 writing behaviors and linguistic complexity. *Language Learning.*

Richard, J., Platt, J., & Weber, H. (1985). *Longman dictionary of applied linguistics*, London, UK: Longman.

Roberts, L. (2012). Review article: Psycholinguistic techniques and resources in second language acquisition research. *Second Language Research, 28*(1), 113-127.

Roberts, L., & Siyanova-Chanturia, A. (2013). Using eye-tracking to investigate topics in L2 acquisition and L2 processing. *Studies in Second Language Acquisition, 35,* 213-235.

Robinson, P. (1995a). Aptitude, awareness and the fundamental similarity of implicit and explicit second language learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 303-357). Honolulu, HI: University of Hawai'i Press.

Robinson, P. (1995b). Task complexity and second language narrative discourse. *Language Learning, 45*(1), 99-140.

Robinson, P. (1995c). Attention, memory and the "noticing" hypothesis. *Language Learning, 45*(2), 283-331.

Robinson, P. (2001a). Task complexity, cognitive resources and syllabus design: A triadic framework for examining task influences on SLA. In. P. Robinson

(Ed.), *Cognition and second language instruction* (pp. 193-226). Cambridge, UK: Cambridge University Press.

Robinson, P. (2001b). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*(1), 27-57.

Robinson, P. (2005a). Aptitude and second language acquisition. *Annual Review of Applied Linguistics, 25,* 45-73.

Robinson, P. (2005b). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics, 43*, 1-32.

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics, 45*, 193-213.

Robinson, P. (2011) *Researching second language task complexity: Task demands, language learning and language performance*. Amsterdam, The Netherlands: John Benjamins.

Robinson, P., Mackey, A., Gass, S. M., & Schmidt, R. (2012). Attention and awareness in second language acquisition. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 247-267). London, UK: Routledge.

Rogers, J. R. (2016). *Developing implicit and explicit knowledge of L2 case marking under incidental learning conditions*. Unpublished dissertation, University College London Institute of Education, London, UK.

Rosa, E., & Leow, R. P. (2004). Awareness, different language conditions, and second language development. *Applied Psycholinguistics*, 25, 269-292.

Rosa, E., & O'Neill, M. D. (1999). Explicitness, intake, and the issue of awareness. Another piece to the puzzle. *Studies in Second Language Acquisition, 21*(4), 511-556.

Rott, S. (2005). Processing glosses: A qualitative exploration of how form-meaning connections are established and strengthened. *Reading in a Foreign Language, 17*(2), 95-124.

Rott, S., & Williams, J. (2003). Making form-meaning connections while reading: A qualitative analysis of word processing. *Reading in a Foreign Language, 15*(1), 45-75.

Rott, S., Williams, J., & Cameron, R. (2002). The effect of multiple-choice L1 glosses and input-output cycles on lexical acquisition and retention. *Language Teaching Research, 6,* 183-222.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, *23*(4), 441–474.

Sagarra, N. (2007). From CALL to face-to-face interaction: The effect of computer-delivered recasts and working memory on L2 development. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 229-248). Oxford: Oxford University Press.

Sagarra, N. (2008). Working memory and L2 processing of redundant grammatical forms. In Z. Han (Ed.), *Understanding second language process* (pp. 133-147). Clevedon, UK: Multilingual Matters.

Sagarra, N., & Ellis, N. C. (2013). From seeing adverbs to seeing verbal morphology: Language experience and adult acquisition of L2 tense. *Studies in Second Language Acquisition 35*, 261-290.

Sagarra, N., & Abbuhl, R. (2013). Optimizing the noticing of recasts via computer-delivered feedback: Evidence that oral input enhancement and working memory help second language learning. *Modern Language Journal, 97*(1), 196-216.

Sasayama, S. (2016). Is a 'complex' task really complex? Validating the assumption of cognitive task complexity. *Modern Language Journal, 100*(1), 231-254.

Sasayama, S., Malicka, A., & Norris, J. (2015). Primary challenges in cognitive task complexity research: Results of a comprehensive research synthesis. Paper presented at the 6th Biennial International Conference on Task-Based Language Teaching (TBLT), Leuven, Belgium.

Sawyer, M., & Ranta, L. (2001). Aptitude, individual differences and instructional design. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 319-353). New York: Cambridge University Press.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*, 129-158.

Schmidt, R. (1995). Consciousness and foreign language learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 1-63). Honolulu, HI: University of Hawai'i at Manoa: Second Language Teaching and Curriculum Center.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). Cambridge, UK: Cambridge University Press.

Serafini, E. J., & Sanz, C. (2015). Evidence for the decreasing impact of cognitive ability on second language development as proficiency increases. *Studies in Second Language Acquisition, 37,* 1-40.

Service, E. (1992). Phonology, working memory, and foreign-language learning. *Quarterly Journal of Experimental Psychology*, *45A*, 21–50.

Service, E., & Kohonen, V. (1995). Is the relation between phonological memory and foreign language learning accounted for by vocabulary acquisition? *Applied Psycholinguistics*, *16*, 155–172.

Sharwood Smith, M. (1986). Comprehension versus acquisition: Two ways of processing input. *Applied Linguistics, 7*(3), 239-256.

Sharwood Smith, M. (1991). Speaking to many minds: On the relevance of different types of language information for the L2 learner. *Second Language Research, 72*, 118-132.

Sharwood Smith, M. (1993). Input enhancement in instructed SLA: Theoretical bases. *Studies in Second Language Acquisition, 15*, 165-179.

Shin, J. A. (2011). Overpassivization errors in Korean college students' English writings. *Korean Journal of Applied Linguistics, 27*(3), 255-273.

Shiotsu, T. (2009). Reading ability and components of word recognition speed: The case of L1-Japanese EFL learners. In Z.-H. Han & N. J. Anderson (Eds.), *Second language reading research and instruction: Crossing the boundaries* (pp. 15-39). Ann Arbor, MI: The University of Michigan Press.

Shook, D. J. (1994). FL/L2 reading, grammatical information, and the input to intake phenomenon. *Applied Language Learning, 5*, 57-93.

Siyanova-Chanturia, A., & Conklin, K., Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research, 27*, 1-22.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics, 17*(1), 38-62.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2002). A non-marginal role for tasks. *ELT Journal, 56*(3), 289-295.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510-532.

Skehan, P. (2014). *Processing perspectives on task performance*. London, England: John Benjamins.

Skehan, P. & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.), *Cognition and second language learning* (pp. 183–205). New York: Cambridge University Press.

Skehan, P., Xiaoyue, B., Qian, L., & Wang, Z. (2012). The task is not enough: Processing approaches to task-based performance. *Language Teaching Research, 16*(2), 170-187.

Smith, F. (1973). *Psycholinguistics and reading.* New York: Holt, Rinehart & Winston.

Smith E. E., & Jonides, J. (1997). Storage and executive processes in the frontal lobes. *Science, 283*, 1657-1661.

Sonbul, S. and Schmitt, N. (2013). Explicit and implicit lexical knowledge: Acquisition of collocations under different input conditions. *Language Learning, 63*(1), 121–159.

Sorace, A. (2000). Gradients in auxiliary selection with intransitive verbs. *Language, 76*, 859-890.

Speciale, G., Ellis, N. C., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics, 25*, 293-432.

Spinner, P., Gass, S. M., & Behney, J. (2013). Ecological validity in eye-tracking. *Studies in Second Language Acquisition, 35*, 389-415.

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly, 16*, 32-71.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235-253). Rowley, MA: Newbury house.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honor of H. G. Widdowson* (pp. 125-144). Oxford, UK: Oxford University Press.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook on research in second language teaching and learning* (pp. 471-83). Mahwah, NJ: Lawrence Erlbaum Associates.

Swain, M., & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *The Modern Language Journal, 82*, 320-337.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251-296.

Taillefer, G. E. (1996). L2 reading ability: Further insight into the short-circuit hypothesis. *Modern Language Journal, 80*, 461-477.

Tavakoli, P. (2009). Investigating task difficulty: learners' and teachers' perceptions. *International Journal of Applied linguistics, 19*(1), 1-25.

Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning, 61*(1), 37-72.

Thomas, E. A. C., & Weaver, W. B. (1975). Cognitive processing and time perception. *Perception and Psychophysics, 17,* 363-367.

Tobii Technology. (n.d.). *Tobii Studio 2.2* [Eye-tracking software]. Stockholm, Sweden.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Cambridge, MA: Harvard University Press.

Tomlin, R. S., & Villa, V. (1994). Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition, 16*, 183-203.

Trofimovich, P., Ammar, A., & Gatbonton, E. (2007). How effective are recasts? The role of attention, memory, and analytic ability. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 171-195). Oxford: Oxford University Press.

Tunmer, W., & Hoover, W. (1992). Cognitive and linguistic factors in learning to read. In P. Gough, L. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 175-214). Hillsdale, NJ: Erlbaum.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28*, 127-154.

Unsworth, N., Heitz, R., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*(3), 498-505.

Urquhart, A. H., & Weir, C. J. (1998). *Reading in a second language: Process, product, and practice*. New York: Longman.

Van den Branden, K., Bygate, M., & Norris, J. (2009). *Task-based language teaching: A reader*. Amsterdam, The Netherlands: John Benjamins Publishing.

Van der Meer, T., te Grotenhuis, M., & Pelzer, B. (2010). Influential cases in multilevel modeling: A methodological comment. *American Sociological Review, 75,* 173-178.

Van Gerven, P. W. M., Paas, F. G. W. C., van Merriënboer, J. J. G., & Schmidt, H. G. (2002). Cognitive load theory and aging: effects of worked examples on training efficiency. *Learning and Instruction, 12*, 87-105.

VanPatten, B. (1990). Attending to form and content in the input. *Studies in Second Language Acquisition, 12*, 287-301.

VanPatten, B. (1996). *Input processing and grammar instruction in second language acquisition*. Norwood, NJ: Ablex Publishing Corporation.

VanPatten, B. (2004). *Processing instruction: Theory, research and commentary.* Mahwah, NJ: Lawrence Erlbaum Associates.

VanPatten, B. (2012). Input processing. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 268-281). London, UK: Routledge.

VanPatten, B., & Cadierno, T. (1993). Explicit instruction and input processing. *Studies in Second Language Acquisition, 15*, 225-243.

Verhagen, J., Leseman, P., & Messer, M. (2015). Phonological memory and the acquisition of grammar in child L2 learners. *Language Learning, 65*(2), 417-448.

Walter, C. (2004). Transfer of reading comprehension skills to L2 is linked to mental representations to text and to L2 working memory. *Applied Linguistics, 25*(3), 315-339.

Watanabe, Y. (1997). Input, intake, and retention: Effects of increased processing on incidental learning of foreign language vocabulary. *Studies in Second Language Acquisition, 19*, 289-307.

Waters, G. S., & Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *The Quarterly Journal of Experimental Psychology, 49A*(1), 51-79.

Wen, Z. (2012). Working memory and second language learning. *International Journal of Applied Linguistics, 22*(1), 1-22.

White, J. (1998). Getting the learner's attention. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition (*pp. 85-113). Cambridge, UK: Cambridge University Press.

White, L. (1989). *Universal grammar and second language acquisition*. Amsterdam, The Netherlands: John Benjamins.

White, L., & Juffs, A. (1998). Constraints on wh-movement in two different contexts of nonnative language acquisition: competence and processing. In S. Flynn, G. Martohardjono, & W. O'Neil (Eds.), *The generative study of second language acquisition* (pp. 111-129). Mahwah, NJ. L. Erlbaum.

White, R.V. (1988). *The ELT Curriculum*. Oxford: Blackwell.Wickens, C. D. (1992). *Engineering psychology and human performance.* New York, NY: Harper Collins.

Wickens, C. D. (2007). Attention to the second language. *IRAL, 45*, 177-191.

Wilkins, D. (1976). *Notional syllabuses*. Hove: Language Teaching Publications.

Williams, J. N. (1999). Memory, attention, and inductive learning. *Studies in Second Language Acquisition, 21*, 1-48.

Winke, P. (2013). The effects of input enhancement on grammar learning and comprehension: a modified replication of Lee (2007) with eye-movement data. *Studies in Second Language Acquisition, 35*, 323-352.

Winke, P., Gass, S., & Sydorenko, T. (2013). Factors influencing the use of captions by foreign language learners: An eye-tracking study. *Modern Language Journal, 97*(1), 254-275.

Winke, P., Godfroid, A., & Gass, S. (2013). Introduction to the special issue: Eye-movement recordings in second language acquisition research. *Studies in Second Language Acquisition, 35*(2), 205-212.

Winter, B. (2013). *Linear models and linear mixed effects models in R with linguistic applications*. arXiv:1308.5499. [http://arxiv.org/pdf/1308.5499.pdf]

Wong, W. (2003). The effects of textual enhancement and simplified input on L2 comprehension and acquisition of non-meaningful grammatical form. *Applied Language Learning, 14*, 109-132.

Yamashita, J. (2003). Processes of taking a gap-filling test: Comparison of skilled and less skilled EFL readers. *Language Testing*, *20*(3), 267–293.

Yano, Y., Long, M. H., & Ross, S. (1994). The effects of simplified and elaborated texts on foreign language comprehension. *Language Learning, 44*, 189-219.

Yilmaz, Y. (2011). Task effects on focus on form in synchronous computer-mediated communication. *Modern Language Journal, 95*(1), 115-132.

Yilmaz, Y. (2013). Relative effects of explicit and implicit feedback: The role of working memory capacity and language analytic ability. *Applied Linguistics, 34*(3), 344-368.

Yoshimura, F. (2006). Does manipulating foreknowledge of output tasks lead to differences in reading behavior, text comprehension and noticing of language form? *Language Teaching Research, 10*(4), 419-434.

Young, D. J. (1999). Linguistic simplification of SL reading material: Effective instructional practice? *Modern Language Journal, 83*(3), 350-366.

Zobl, H. (1989). Canonical typological structures and ergativity in English L2 acquisition. In S. M. Gass & J. Schachter (Eds.), *Linguistic perspectives on second language* (pp. 203–221). Cambridge University Press.

Zyzik, E. (2009). The role of input revisited: Nativist versus usage-based models. *L2 Journal, 1*, 42–61.

Zyzik, E., & Polio, C. (2008). Epilogue: A tale of two copulas. *Bilingualism: Language and Cognition*, *11*(3), 383-385.

APPENDICES

Appendix A-1. Information sheet and consent form for Study 1

STUDY INFORMATION SHEET:

Second language reading comprehension and second language proficiency

I am a doctoral student at the Institute of Education, University of London, interested in second language reading and learning. I would like to invite you to participate in a research study that examines the relationship between second language reading comprehension and second language proficiency.

If you decide to participate, I will ask you to take part in four 1-hour sessions in a computer laboratory at Sogang University. In the 1st session, you will be asked to complete a grammar test and an English proficiency test. In the 2nd session, you will be asked to perform a reading comprehension task and a memory test. In the 3rd session, you will be asked to perform another reading comprehension task and a grammar and vocabulary test. In the 4th session, you will be asked to complete memory tests and another grammar and vocabulary test. In addition, you will be asked to complete an oral memory test, which takes about 10 minutes, individually with the researcher.

In return for your participation, I am able to offer you a *KRW 40,000 Starbucks gift card* at the end of the 4th session. As soon as the data analysis is completed, I will also share the overall results of the study with you.

Any data obtained from you will be kept securely. At every stage of the project and beyond, your name will remain confidential. Your identity will be anonymised by the use of a unique identifier. The overall results of the study will be used for my doctoral research project and not be shared with others. The results will also be presented at professional conferences and in research publications.

You are free to withdraw from the study at any time without reason and without any impact on you. If you decide to withdraw, any data collected from you will be destroyed. If you have any queries about the study, please feel free to contact Jookyoung Jung at jookyoung.jung@gmail.com, 010-2584-5170.

I would be very grateful if you would agree to take part!

Jookyoung Jung
PhD student at the Institute of Education, University of London

CONSENT FORM

Project title: ***Second language reading comprehension and second language proficiency***

		YES	NO
1.	I have read and had explained to me by Jookyoung Jung the Information Sheet relating to this project.	□	□
2.	I have had explained to me the purposes of the project and what will be required of me, and any questions have been answered to my satisfaction. I agree to the arrangements for my participation as described in the Information Sheet.	□	□
3.	I understand that my participation is entirely voluntary and that I have the right to withdraw from the project any time.	□	□
4.	I agree with the contents of this Consent Form and have received the accompanying Information Sheet.	□	□

Name:

Signed:

Date:

Appendix A-2. Information sheet and consent form for Study 2

STUDY INFORMATION SHEET:

Second language reading comprehension and second language proficiency

I am a doctoral student at the Institute of Education, University of London, interested in second language reading and learning. I would like to invite you to participate in a research study that examines the relationship between second language reading comprehension and second language proficiency.

If you decide to participate, I will ask you to take part in four 1-hour sessions in a computer laboratory at Korea University. In the 1st session, you will be asked to complete a grammar test and an English proficiency test. In the 2nd session, you will be asked to perform a reading comprehension task and a memory test. In the 3rd session, you will be asked to perform another reading comprehension task and a grammar and vocabulary test. In the 4th session, you will be asked to complete memory tests and another grammar and vocabulary test. In addition, you will be asked to complete an oral memory test, which takes about 10 minutes, individually with the researcher.

In return for your participation, I am able to offer you *KRW 40,000* at the end of the 4th session. As soon as the data analysis is completed, I will also share the overall results of the study with you.

Any data obtained from you will be kept securely. At every stage of the project and beyond, your name will remain confidential. Your identity will be anonymised by the use of a unique identifier. The overall results of the study will be used for my doctoral research project and not be shared with others. The results will also be presented at professional conferences and in research publications.

You are free to withdraw from the study at any time without reason and without any impact on you. If you decide to withdraw, any data collected from you will be destroyed. If you have any queries about the study, please feel free to contact Jookyoung Jung at jookyoung.jung@gmail.com, 010-2584-5170.

I would be very grateful if you would agree to take part!

Jookyoung Jung
PhD student at the Institute of Education, University of Lond

CONSENT FORM

Project title: ***Second language reading comprehension and second language proficiency***

	YES	NO
1. I have read and had explained to me by Jookyoung Jung the Information Sheet relating to this project.	□	□
2. I have had explained to me the purposes of the project and what will be required of me, and any questions have been answered to my satisfaction. I agree to the arrangements for my participation as described in the Information Sheet.	□	□
3. I understand that my participation is entirely voluntary and that I have the right to withdraw from the project any time.	□	□
4. I agree with the contents of this Consent Form and have received the accompanying Information Sheet.	□	□

Name:

Signed:

Date:

Appendix A-3. Information sheets and consent form for Study 3

STUDY INFORMATION SHEET:

Second language reading comprehension and second language proficiency

I am a doctoral student at the Institute of Education, University of London, interested in second language reading and learning. I would like to invite you to participate in a research study that examines the relationship between second language reading comprehension and second language proficiency.

If you decide to participate, I will ask you to take part in a three-hour long session in a computer laboratory at the Institute of Education. You will be asked to complete an English proficiency test and perform two reading comprehension tasks. After finishing the reading comprehension tasks, you will be asked to take part in an interview to share your reflections on the task performance.

In return for your participation, I am able to offer you £25 at the end of the session. As soon as the data analysis is completed, I will also share the overall results of the study with you.
Any data obtained from you will be kept securely. At every stage of the project and beyond, your name will remain confidential. Your identity will be anonymised by the use of a unique identifier. The overall results of the study will be used for my doctoral research project and not be shared with others. The results will also be presented at professional conferences and in research publications.

You are free to withdraw from the study at any time without reason and without any impact on you. If you decide to withdraw, any data collected from you will be destroyed. If you have any queries about the study, please feel free to contact Jookyoung Jung at jookyoung.jung@gmail.com.

I would be very grateful if you would agree to take part!

Jookyoung Jung
PhD student at the Institute of Education, University of London

STUDY INFORMATION SHEET:

Second language reading comprehension and second language proficiency

I am a doctoral student at the Institute of Education, University of London, interested in second language reading and learning. I would like to invite you to participate in a research study that examines the relationship between second language reading comprehension and second language proficiency.

If you decide to participate, I will ask you to take part in a two-hour long session in a computer laboratory at the Institute of Education. You will be asked to complete an English proficiency test and perform two reading comprehension tasks.

In return for your participation, I am able to offer you £15 at the end of the session. As soon as the data analysis is completed, I will also share the overall results of the study with you.
Any data obtained from you will be kept securely. At every stage of the project and beyond, your name will remain confidential. Your identity will be anonymised by the use of a unique identifier. The overall results of the study will be used for my doctoral research project and not be shared with others. The results will also be presented at professional conferences and in research publications.

You are free to withdraw from the study at any time without reason and without any impact on you. If you decide to withdraw, any data collected from you will be destroyed. If you have any queries about the study, please feel free to contact Jookyoung Jung at jookyoung.jung@gmail.com.

I would be very grateful if you would agree to take part!

Jookyoung Jung
PhD student at the Institute of Education, University of London

CONSENT FORM

Project title: ***Second language reading comprehension and second language proficiency***

	YES	NO
1. I have read and had explained to me by Jookyoung Jung the Information Sheet relating to this project.	□	□
2. I have had explained to me the purposes of the project and what will be required of me, and any questions have been answered to my satisfaction. I agree to the arrangements for my participation as described in the Information Sheet.	□	□
3. I understand that my participation is entirely voluntary and that I have the right to withdraw from the project any time.	□	□
4. I agree with the contents of this Consent Form and have received the accompanying Information Sheet.	□	□

Name:

Signed:

Date:

Appendix B-1. Grammaticality judgment sentences for Study 1

[Target unaccusative verbs]

1. Plastic is not decomposed easily.
2. The dead body began to decompose.
3. After the storm, snow was accumulated on the sidewalk.
4. After the storm, leaves accumulated around the tall trees.
5. Fallen leaves were drifted in the air.
6. The dumped boat drifted across the sea.
7. Flowers are soft, so they are not fossilized easily.
8. Snake bones are so soft and do not fossilize easily.
9. His class is consisted of group works.
10. The crew consists of five experts.
11. The tension disappeared completely.
12. The sun was disappeared completely.
13. The storm has been subsided.
14. The back pain hasn't subsided.
15. The soldiers ceased all dangerous actions at once.
16. The job will be ceased to exist on his retirement.
17. Rain water collected in the old tank very slowly.
18. Micro-organisms are collected on the seafloor.
19. Homo sapiens were originated in Africa.
20. Potatoes originated in South America.
21. Some old churches in Europe date to the 4th century.
22. The debate among researchers is dated back to 1986.
23. The discussion has been persisted for more than two hours.
24. The boy's high fever has persisted for more than two days.
25. The Gothic style evolved from the Romanesque style.
26. Domestic dogs were evolved from European wolves.
27. Korean pop emerged as a global trend.
28. A shiny star was emerged from the cloud.
29. The oil price ascended about 5 percent last year.
30. The salmon was ascended to the river to spawn.
31. The ozone layer has been diminished.
32. His influence has never diminished.
33. The Dutch and Spanish also settled here.
34. One day my wife wants to be settled here.

[Novel unaccusative verbs]

35. My dad was remained quiet the whole time.
36. Her husband remained in the room quietly.
37. All the windows broke after the last windstorm.
38. The waves were broken when they reached the beach.
39. The traffic lights changed from green to red.
40. Suddenly the rain was changed into showers.
41. All the colored leaves fell in the windstorm last night.
42. The temperature was fallen by 10 degrees last night.
43. Human skin is burned at around 54 degrees.
44. Natural gas burns at about 3000 degrees.
45. The baby's toy ran out of batteries and was stopped.
46. It rained heavily for a while and the subways stopped.
47. She was appeared to be shocked very much.

48. He may appear smaller than his real height.
49. The terrible crime occurred near the town.
50. The latest crime was occurred at midnight.

[Distracters]

51. If I studied a little harder, I should have passed the exam.
52. He is rather an old man.
53. Heating and cooling can cause matter to expand and contract.
54. There is nothing to fear, provided that you are just.
55. Our product is superior to our competitor's.
56. I found the missing envelope just outside of the house.
57. I am looking forward to see you soon.
58. The play was so bored that I fell asleep.
59. Comparing with his brother, he is not so clever.
60. I have two brothers; one lives in Seoul and another in Busan.
61. She looks very fatter than she was a month before.
62. This chair, that has been broken for weeks, must be repaired.
63. I prefer coffee than tea.
64. My father is usually going to his office by bus.
65. The teacher suggested to us that we study English very hard.
66. Only after the next morning I knew the fact.
67. You had better not going out in such a heavy snowfall.
68. Industrial diamonds used to be cut hard materials.
69. The tropical forests have been destroyed since the past fifty years.
70. If you will not do so, neither I will.
71. Anybody in this class does not deserve to be happy.
72. She was doing whichever she could to stay alive.
73. He is a talented and imaginable writer.
74. Barely no one noticed that something was wrong.
75. We made him to write a letter of apology.
76. The bathroom walls are with marble and tiles.
77. I couldn't help noticing that you weren't very polite.
78. Be careful lest you should hit your head against the post.
79. Taking all things into consideration, he cannot be the criminal.
80. Had you not helped me, I would have failed.

Appendix B-2. Grammaticality judgment sentences for Study 2 and Study 3

[Target unaccusative verbs]

1. CO_2 in the air has been accumulated quickly.
2. The country's wealth has accumulated quickly.
3. Dark clouds were drifted in the air.
4. The old boat drifted out to the sea.
5. His class is consisted of group works.
6. The crew consisted of five experts.
7. The sun was soon disappeared.
8. The tension soon disappeared.
9. Heavy materials are settled to the bottom faster.
10. The pollen in the honey will settle to the bottom.
11. The island was subsided about six inches.
12. The island subsided after the earthquake.
13. Brain death is when all brain activities are ceased.
14. The young soldiers ceased all violent actions at once.
15. Water was collected on the ground after the rain.
16. Rain water collected in the large tank very fast.
17. Homo sapiens were originated in Africa.
18. Potatoes originated in South America.
19. His interest in flying is dated to his childhood.
20. The history of music dates to ancient times.
21. The problem has been persisted for years.
22. The boy's high fever has persisted for days.
23. Domestic dogs were evolved from European wolves.
24. The Gothic style evolved from the Romanesque style.
25. A shiny star was emerged from the cloud.
26. The black bird slowly emerged from the fog.
27. The balloon was ascended high up in the sky.
28. Our new car ascended the road very easily.
29. Rain chances will be diminished.
30. The supply of oil will diminish.

[Novel unaccusative verbs]

31. Smoke was appeared on the horizon.
32. This story appears in many paintings.
33. My dad was remained quiet the whole time.
34. My husband silently remained in the room.
35. The wooden bridge was broken suddenly.
36. The rope finally broke with a loud sound.
37. Suddenly the rain was changed into showers.
38. The traffic lights changed from green to red.
39. The temperature was fallen by 10 degrees last night.
40. All the colored leaves fell during the windstorm last night.
41. Human skin is burned at around 54 degrees.
42. Natural gas burns at around 3000 degrees.
43. The rain was finally stopped.
44. His laughter stopped suddenly.
45. The latest crime was occurred at midnight.
46. The terrible crime occurred near the town.

[Distracters]

47. Be careful lest you should hit your head against the post.
48. Taking all things into consideration, he cannot be the criminal.
49. Had you not helped me, I would have failed.
50. If I studied a little harder, I should pass the exam.
51. He is rather an old man.
52. Heating and cooling can cause matter to expand and contract.
53. There is nothing to fear, provided that you are just.
54. Our product is superior to our competitor's.
55. I found the missing envelope just outside of the house.
56. I couldn't help noticing that you weren't very polite.
57. The bathroom walls are in marble and tiles.
58. If you will not do so, neither will I.
59. I am looking forward to seeing you soon.
60. Compared with his brother, he is not so clever.
61. She looks much fatter than she was a month ago.
62. Some fleas have one or two eyes, but others have none.
63. If I were to have no friends, I would be terribly lonely.
64. The play was so bored that I fell asleep.
65. Dogs rely more on their sense of smell than for any other senses.
66. I have two brothers; one lives in Seoul and another in Pusan.
67. This chair, that has been broken for weeks, must be repaired.
68. I prefer coffee than tea.
69. My father is usually going to his office by bus.
70. The teacher suggested to us that we study English very hard.
71. Only after the next morning I knew the fact.
72. You had better not going out in such a heavy snowfall.
73. Industrial diamonds used to be cut hard materials.
74. The tropical forests have been destroyed since the past fifty years.
75. Anybody in this class does not deserve to be happy.
76. She was doing whichever she could to stay alive.
77. He is a talented and imaginable writer.
78. Barely no one noticed that something was wrong.
79. We made him to write a letter of apology.
80. Despite it was weekend, I went to work as usual.

Appendix C-1. Vocabulary form recognition test for Study 1

지문: 두번의 리딩 텍스트에서 다음의 단어를 본 기억이 나면 "Yes"를, 기억이 나지 않으면 "No"를 선택해 주세요.
Directions: If you remember seeing the words below from the reading texts, check "Yes." It not, check "No."

1. phanlin	□ Yes	□ No	**11. fration**	□ Yes	□ No
2. writchy	□ Yes	□ No	**12. zenters**	□ Yes	□ No
3. vidoses	□ Yes	□ No	**13. golands**	□ Yes	□ No
4. scaders	□ Yes	□ No	**14. stragon**	□ Yes	□ No
5. cainmat	□ Yes	□ No	**15. morbits**	□ Yes	□ No
6. liphore	□ Yes	□ No	**16. tralion**	□ Yes	□ No
7. flarris	□ Yes	□ No	**17. klaners**	□ Yes	□ No
8. bolloug	□ Yes	□ No	**18. phosens**	□ Yes	□ No
9. dasters	□ Yes	□ No	**19. stovons**	□ Yes	□ No
10. lurgled	□ Yes	□ No	**20. cabrons**	□ Yes	□ No

Appendix C-2. Vocabulary form recognition test for Study 2

지문: 두번의 리딩 텍스트에서 다음의 단어를 본 기억이 나면 "Yes"를, 기억이 나지 않으면 "No"를 선택해 주세요.
Directions: If you remember seeing the words below from the reading texts, check "Yes." It not, check "No."

1. phanlin	□ Yes	□ No	**11. fration**	□ Yes	□ No
2. writchy	□ Yes	□ No	**12. zenters**	□ Yes	□ No
3. vidoses	□ Yes	□ No	**13. golands**	□ Yes	□ No
4. scaders	□ Yes	□ No	**14. stragon**	□ Yes	□ No
5. cainmat	□ Yes	□ No	**15. morbits**	□ Yes	□ No
6. liphore	□ Yes	□ No	**16. tralion**	□ Yes	□ No
7. flarris	□ Yes	□ No	**17. klenear**	□ Yes	□ No
8. bolloug	□ Yes	□ No	**18. phosens**	□ Yes	□ No
9. dasters	□ Yes	□ No	**19. stovons**	□ Yes	□ No
10. lurgled	□ Yes	□ No	**20. cabrons**	□ Yes	□ No

Appendix D-1. Vocabulary meaning recognition test for Study 1

지문: 다음의 각 단어 의미와 가장 가까운 것을 고르세요. 추측하지 마세요.
Directions: Choose the meaning of each of the words below. Do not try to guess.

1. fration
(1) 발견
(2) 변화
(3) 부재
(4) 잘 모르겠다.

2. phanlin
(1) 변화
(2) 부재
(3) 발견
(4) 잘 모르겠다.

3. zenters
(1) 후손
(2) 힌트
(3) 공원
(4) 잘 모르겠다.

4. writchy
(1) 힌트
(2) 공원
(3) 후손
(4) 잘 모르겠다.

5. golands
(1) 분출
(2) 후손
(3) 발견
(4) 잘 모르겠다.

6. vidoses
(1) 발견
(2) 후손
(3) 분출
(4) 잘 모르겠다.

7. stragon
(1) 해변
(2) 공원
(3) 바닥
(4) 잘 모르겠다.

8. scaders
(1) 공원
(2) 바닥
(3) 해변
(4) 잘 모르겠다.

9. morbits
(1) 힌트
(2) 후손
(3) 바닥
(4) 잘 모르겠다.

10. cainmat
(1) 후손
(2) 바닥
(3) 힌트
(4) 잘 모르겠다.

11. tralion
(1) 천적
(2) 변화
(3) 해변
(4) 잘 모르겠다.

12. liphore
(1) 변화
(2) 해변
(3) 천적
(4) 잘 모르겠다.

13. klaners
(1) 천적
(2) 바닥
(3) 공원
(4) 잘 모르겠다.

14. flarris
(1) 바닥
(2) 공원
(3) 천적
(4) 잘 모르겠다.

15. phosens
(1) 분출
(2) 발견
(3) 부재
(4) 잘 모르겠다.

16. bolloug
(1) 발견
(2) 부재
(3) 분출
(4) 잘 모르겠다.

17. stovons
(1) 해변
(2) 힌트
(3) 천적
(4) 잘 모르겠다.

18. dasters
(1) 힌트
(2) 천적
(3) 해변
(4) 잘 모르겠다.

19. cabrons
(1) 분출
(2) 부재
(3) 변화
(4) 잘 모르겠다.

20. lurgled
(1) 부재
(2) 변화
(3) 분출
(4) 잘 모르겠다.

Appendix D-2. Vocabulary meaning recognition test for Study 2

지문: 다음의 각 단어 의미와 가장 가까운 것을 고르세요. 추측하지 마세요.
Directions: Choose the closest meaning of each of the words below. Do not try to guess.

1. fration
(1) 발견
(2) 변화
(3) 부재
(4) 잘 모르겠다.

2. phanlin
(1) 변화
(2) 부재
(3) 발견
(4) 잘 모르겠다.

3. zenters
(1) 후손
(2) 힌트
(3) 표면
(4) 잘 모르겠다.

4. writchy
(1) 힌트
(2) 표면
(3) 후손
(4) 잘 모르겠다.

5. golands
(1) 분출
(2) 후손
(3) 발견
(4) 잘 모르겠다.

6. vidoses
(1) 발견
(2) 후손
(3) 분출
(4) 잘 모르겠다.

7. stragon
(1) 해변
(2) 표면
(3) 바닥
(4) 잘 모르겠다.

8. scaders
(1) 표면
(2) 바닥
(3) 해변
(4) 잘 모르겠다.

9. morbits
(1) 힌트
(2) 후손
(3) 바닥
(4) 잘 모르겠다.

10. cainmat
(1) 후손
(2) 바닥
(3) 힌트
(4) 잘 모르겠다.

11. tralion
(1) 조건
(2) 변화
(3) 해수
(4) 잘 모르겠다.

12. liphore
(1) 변화
(2) 해수
(3) 조건
(4) 잘 모르겠다.

13. klenear
(1) 조건
(2) 바닥
(3) 표면
(4) 잘 모르겠다.

14. flarris
(1) 바닥
(2) 표면
(3) 조건
(4) 잘 모르겠다.

15. phosens
(1) 분출
(2) 발견
(3) 부재
(4) 잘 모르겠다.

16. bolloug
(1) 발견
(2) 부재
(3) 분출
(4) 잘 모르겠다.

17. stovons
(1) 해수
(2) 힌트
(3) 조건
(4) 잘 모르겠다.

18. dasters
(1) 힌트
(2) 조건
(3) 해수
(4) 잘 모르겠다.

19. cabrons
(1) 분출
(2) 부재
(3) 변화
(4) 잘 모르겠다.

20. lurgled
(1) 부재
(2) 변화
(3) 분출
(4) 잘 모르겠다.

Appendix E. Questionnaires

연구참여 전 설문조사

*** 주어진 빈칸 (_____) 에 답을 적거나 □을 체크해주세요.**

1. 당신은: □ 여성 □ 남성

2. 전공이 무엇입니까?: _____

3. 몇 살 이십니까?: _____ 세.

4. 언제 처음으로 영어를 배우기 시작했습니까?: _____ 세.

5. 영어를 사용하는 나라에 살거나 공부한 경험이 있습니까?: □ 네 (5-1 로 가세요.) □ 아니오

   5-1. 나라: _____ 언제: _____

   지낸 기간: _____

6. 한국어와 영어 외에 사용할 수 있는 언어가 있습니까?: □ 네 (6-1 로 가세요.) □ 아니오

   6-1. 어떤 언어입니까?: _____

   그 언어의 수준은 어떻습니까?: □ 초급 □ 중급 □ 상급

   6-2. 그 언어로 읽기도 합니까?: □ 네 □ 아니오

7. TOEFL 점수가 있다면 적어주세요: _____

Background Questionnaire

Directions: Write down your answers in the provided blanks (_____) or check □ that corresponds to your answer.

1. You are: □ Female □ Male

2. What is your major?: _____

3. How old are you?: _____ years old.

4. When did you first begin learning English?: _____ years old.

5. Have you ever stayed/lived in an English-spoken country?: □ Yes (*go to* 5-1.) □ No

   5-1. Where: _____ When: _____

   How long: _____

6. Do you know any language other than English and Korean?:

   □ Yes (*go to* 6-1.) □ No

6-1. What is the language?: _____

Proficiency level?: □ beginner □ intermediate □ advanced

6-2. Do you read in that language?: □ Yes □ No

7. If you have a TOEFL score, please provide it here: _____

<u>과업 후 설문조사</u>

*** 주어진 빈칸 (_____) 에 답을 적거나 □을 체크해주세요.**

내가 **이 과업을 마치는데에는** 총 _____ **분이 걸렸다**.

1. **나는 이 과업이 어려웠다고 생각한다**.

매우 그렇다 그렇다 약간 그렇다 그저그렇다 별로 그렇지않다 그렇지않다 전혀 그렇지않다
□ □ □ □ □ □ □

2. **나는 이 과업을 하면서 좌절감을 느꼈다**.

매우 그렇다 그렇다 약간 그렇다 그저그렇다 별로 그렇지않다 그렇지않다 전혀 그렇지않다
□ □ □ □ □ □ □

3. **나는 이 과업을 잘 못한 것 같다**.

매우 그렇다 그렇다 약간 그렇다 그저그렇다 별로 그렇지않다 그렇지않다 전혀 그렇지않다
□ □ □ □ □ □ □

4. **나는 이 과업이 흥미롭다고 느꼈다**.

매우 그렇다 그렇다 약간 그렇다 그저그렇다 별로 그렇지않다 그렇지않다 전혀 그렇지않다
□ □ □ □ □ □ □

5. **나는 이런 과업을 또 해보고 싶다**.

매우 그렇다 그렇다 약간 그렇다 그저그렇다 별로 그렇지않다 그렇지않다 전혀 그렇지않다
□ □ □ □ □ □ □

6. **나는 이 과업에서 쓰인 지문의 주제가 익숙했다**.

매우 그렇다 그렇다 약간 그렇다 그저그렇다 별로 그렇지않다 그렇지않다 전혀 그렇지않다
□ □ □ □ □ □ □

7. **나는 이 과업을 수행하기 위해 지적인 노력을 많이 기울였다**.

매우 그렇다 그렇다 약간 그렇다 그저그렇다 별로 그렇지않다 그렇지않다 전혀 그렇지않다
□ □ □ □ □ □ □

8. **나는 이 과업을 수행하는 동안 어려움을 겪었다**.

매우 그렇다 그렇다 약간 그렇다 그저그렇다 별로 그렇지않다 그렇지않다 전혀 그렇지않다

□ □ □ □ □ □ □

9. 나는 나의 과업수행 결과에 자신이 있다.

매우 그렇다 그렇다 약간 그렇다 그저그렇다 별로 그렇지않다 그렇지않다 전혀 그렇지않다
□ □ □ □ □ □ □

10. 나는 이 과업이 유익하다고 생각한다.

매우 그렇다 그렇다 약간 그렇다 그저그렇다 별로 그렇지않다 그렇지않다 전혀 그렇지않다
□ □ □ □ □ □ □

11. 이 과업은 내가 긴장을 하게 만들었다.

매우 그렇다 그렇다 약간 그렇다 그저그렇다 별로 그렇지않다 그렇지않다 전혀 그렇지않다
□ □ □ □ □ □ □

12. 나는 이 과업을 즐겼다.

매우 그렇다 그렇다 약간 그렇다 그저그렇다 별로 그렇지않다 그렇지않다 전혀 그렇지않다
□ □ □ □ □ □ □

13. 나는 이 읽기 주제에 대한 배경지식이 있었다.

매우 그렇다 그렇다 약간 그렇다 그저그렇다 별로 그렇지않다 그렇지않다 전혀 그렇지않다
□ □ □ □ □ □ □

14. 나는 이 과업이 힘들다고 생각했다.

매우 그렇다 그렇다 약간 그렇다 그저그렇다 별로 그렇지않다 그렇지않다 전혀 그렇지않다
□ □ □ □ □ □ □

15. 이 과업에 대해 남겨주시고 싶은 말씀이 있다면 아래 빈 칸에

적어주세요:

_____

_____

_____

_____

Post-reading Questionnaire

Directions: Write down your answers in the provided blanks (_____) or check □ that corresponds to your answer.

This task took me _____ minutes to complete.

1. **I thought this task was difficult**.

Strongly agree Agree Slightly agree So-so Slightly disagree Disagree Strongly disagree
☐ ☐ ☐ ☐ ☐ ☐ ☐

2. **I felt frustrated doing this task**.

Strongly agree Agree Slightly agree So-so Slightly disagree Disagree Strongly disagree
☐ ☐ ☐ ☐ ☐ ☐ ☐

3. **I did poorly on this task**.

Strongly agree Agree Slightly agree So-so Slightly disagree Disagree Strongly disagree
☐ ☐ ☐ ☐ ☐ ☐ ☐

4. **I felt this task was interesting**.

Strongly agree Agree Slightly agree So-so Slightly disagree Disagree Strongly disagree
☐ ☐ ☐ ☐ ☐ ☐ ☐

5. **I want to do more tasks like this**.

Strongly agree Agree Slightly agree So-so Slightly disagree Disagree Strongly disagree
☐ ☐ ☐ ☐ ☐ ☐ ☐

6. **I thought the topic of the reading was familiar**.

Strongly agree Agree Slightly agree So-so Slightly disagree Disagree Strongly disagree
☐ ☐ ☐ ☐ ☐ ☐ ☐

7. **I invested a large amount of mental effort to complete this task**.

Strongly agree Agree Slightly agree So-so Slightly disagree Disagree Strongly disagree
☐ ☐ ☐ ☐ ☐ ☐ ☐

8. **I struggled during this task**.

Strongly agree Agree Slightly agree So-so Slightly disagree Disagree Strongly disagree
☐ ☐ ☐ ☐ ☐ ☐ ☐

9. **I am confident about my task performance**.

Strongly agree Agree Slightly agree So-so Slightly disagree Disagree Strongly disagree
☐ ☐ ☐ ☐ ☐ ☐ ☐

10. **I found this task helpful**.

Strongly agree Agree Slightly agree So-so Slightly disagree Disagree Strongly disagree
☐ ☐ ☐ ☐ ☐ ☐ ☐

11. **This task was stressful for me**.

Strongly agree Agree Slightly agree So-so Slightly disagree Disagree Strongly disagree
☐ ☐ ☐ ☐ ☐ ☐ ☐

12. **I enjoyed doing this task**.

Strongly agree Agree Slightly agree So-so Slightly disagree Disagree Strongly disagree
□ □ □ □ □ □ □

13. **I had some background knowledge about the reading topic**.

Strongly agree Agree Slightly agree So-so Slightly disagree Disagree Strongly disagree
□ □ □ □ □ □ □

14. **I thought this task was demanding**.

Strongly agree Agree Slightly agree So-so Slightly disagree Disagree Strongly disagree
□ □ □ □ □ □ □

15. **If you have any comments to the researcher, please provide them in the space below**:

_____

_____

_____

_____

출구 설문조사

* 주어진 빈칸 (_____) 에 생각하시는 대로 적어주세요.

1. **이 연구의 목적이 무엇이라고 생각합니까**?

_____

2. **이 연구를 통해 영어 문법이나 단어를 배운 것이 있습니까**?

_____

3. **이 연구에 참여하는 동안 특정한 영어 문법이나 단어에 주의를 기울였습니까**?

_____

4. 연구 밖에서 영어 문법이나 단어를 찾아보거나 공부했습니까?

_____

5. **이 연구에 참여하면서 남겨주시고 싶은 말씀이 있다면 아래 빈 칸에 적어주세요**:

_____

_____

_____

_____

Exit Questionnaire

Directions: Write down your responses in the provided blanks (_____).

1. **What do you think were the goals of this study**?

_____

2. **Did you learn anything from this study**?

_____

3. **Did you focus on any specific grammatical or lexical feature in this study**?

_____

4. **Did you look up or study any forms or information outside of this study**?

_____

5. **If you have any comments about this task, please provide them in the space below**:

_____

_____

_____

_____

Appendix F. Instruction used for stimulated recall

STIMULATED RECALLS

캠코더의 녹화 버튼을 반드시 누르고 시작한다.
Remember to switch on the recording device at the beginning of the session!

I. Introduction

이제 여러분이 하신 영어독해테스트를 녹화해 놓은 비디오를 보게 되십니다. 이 비디오는 여러분의 눈이 어디를 보고 있었고 어떻게 움직였는지를 보여줍니다. 빨간 원들은 여러분의 눈동자가 어디에 얼만큼 머물렀는지를 보여줍니다. 원이 클수록 더 오래 응시했다는 것을 의미합니다. 원과 원 사이의 선은 여러분의 눈이 어떻게 움직였는지를 나타냅니다.
Now a videotape of your performance on each task will be played again for you. You will be able to see how your eyes moved. Each red circle that you will see represents an eye-gaze. The lines between the circles show how your eyes moved between gazes. The bigger a circle, the longer your eye gaze was.

여러분이 영어독해테스트를 하시면서 생각을 많이/ 열심히 해야 했거나, 어려운 점이 있었거나, 뭔가 말씀하시고 싶은 것이 기억나시면 언제든지 비디오를 멈춰주시고 생각나는 것들을 말씀해주세요. 연구자도 비디오를 멈추고 여러분이 읽으시는 동안 어떤 생각을 하셨는지 몇몇 질문을 할 것입니다.
Y*ou* will be asked to stop the recording whenever you remember that you had to think hard, in other words, when the task was cognitively challenging and mentally effortful. *I* will also stop the recording from time to time and ask you a couple of questions about what you were thinking when you were reading.

Let's do a practice task!

Example Task: stimulated recall 1

연구자는 여러분이 하신 영어독해테스트를 녹화해 놓은 비디오를 보여드릴 겁니다. 저는 여러분이 영어독해테스트를 하던 바로 그 순간에 어떤 생각을 하고 계셨는지에 관심이 있습니다. 저는 여러분이 영어독해테스트를 하는 모습을 봤지만, 그때 여러분이 어떤 생각을 하고 계셨는지는 볼 수가 없었기 때문입니다. 그러므로 제게 여러분께서 독해를 하고 계시던 바로 그 순간에 어떤 생각을 하고 계셨는지를 말씀해 주세요. 여러분께서 지금 생각하시는 것을 말씀하시는 것이 아닙니다.
Again, I'm going to replay the recording of your performance. I am interested in what you were thinking at the time you were reading. I saw you complete the task, but I don't know what you were thinking while you were doing it. So what I'd like you to do is tell me what you were thinking, what was on your mind just then while you were carrying it out. I don't want you to tell me what you think about now.

여러분께서 독해를 하시면서 어려웠던 순간이 기억나시면 비디오를 멈추고 말씀해주세요. 다시 말하면, 정신적으로 노력을 많이 기울여야 했거나 생각을 많이 해야 했던 순간이 있으시면 비디오를 멈춰주세요 (직접 시연해 보인다).
I want you to pause the recording whenever you remember that you had to think hard, in other words, when the task was cognitively challenging and mentally effortful. *(Show the student how to do this).*

저는 비디오를 멈추고 여러분께 몇몇 질문을 할 수도 있습니다. 예를 들면,
I will also interrupt the recording at certain points and ask you a couple of questions.

- What made you ...
- 왜 뒤로 가서 다시 읽으셨나요? read that part again?
- 왜 저 부분을 다시 보셨나요? watch that part again?
- 왜 오래 쳐다보셨나요? look at that part for a long time?

- 왜 단어뜻을 보셨나요? refer to the word meaning?

이해가 가시나요?
Is it clear what we will do?

(학생이 언제든 녹화를 멈추고 발언할 수 있도록 시작한다.)
(Let the student start the recording and pause whenever s/he has a thought to share.)

(학생들이 발언할 때, "아 그렇군요" "좋습니다" "네" 등 이상의 반응을 하지 않는다.
(While the student is voicing his/her thoughts, try not to react to responses other than providing backchannelling cues or non-responses: Oh, mhm, great, good, I see, uh-huh, ok)

(다음 과업에 대해서도 같은 절차를 반복한다.)
(Repeat the same procedure for the rest of the tasks).

Practical tips

- 리콜을 녹음할 때에는 참여자 번호, 텍스트 번호를 먼저 말하고 텍스트마다 따로 녹음한다.
  You might want to tape each task performance separately. This will make it easier to find the relevant recordings.